Using deep learning to investigate neuroanatomical abnormalities in first-episode psychosis

Vieira, Sandra

Awarding institution:
King's College London

Using deep learning to investigate neuroanatomical abnormalities in first-episode psychosis

Sandra Vieira

Institute of Psychiatry, Psychology and Neuroscience King's College London

Thesis submitted to King's College London in fulfilment for the degree of Doctor of Philosophy (PhD)

May 2019

Abstract

Evidence of neuroanatomical abnormalities in subjects with a recent first episode of psychosis (FEP) has been heterogeneous, possibly due to the increased risk of false positives and heterogeneous findings associated with small samples that dominate the literature. In addition, the clinical impact of such findings has been limited. Machine learning promises to overcome this limitation, however, initial attempts to identify FEP have yielded inconsistent results. Within this movement, deep learning has recently emerged as a promising approach in areas such as visual and speech recognition, as well as other areas of medicine. Its ability to capture highly abstract and complex interactions may be useful to capture the characteristic subtle and widespread neuroanatomical changes of FEP.

The overarching aim of this doctoral thesis was to investigate neuroanatomical abnormalities in FEP at the individual level in a mega-analytic approach. Structural Magnetic Resonance Imaging (sMRI) data was collated from five independent studies, totalling 1074 participants. FEP and healthy controls (HC) were first compared using voxel-based morphometry in a large-scale mega-analysis. This was followed by a thorough review of the current evidence for deep learning in psychiatric and neurologic neuroimaging. A deep neural network, along with other well-established methods for comparison, were then used to classify FEP and HC at each site separately to test for the reproducibility of findings. Finally, a deep neural network was used to classify the two groups in a large-scale mega-analysis.

Collectively, results revealed a pattern of fronto-temporal-insular changes identified both at group and individual level. Deep neural networks performed better than traditional machine learning approaches, albeit by a small margin. However, performances were lower than expected overall, ranging between 50 and 70%. Upon interpreting these results, I was able to show evidence for publication bias, suggesting that initial studies may have been over-optimistic. Consist with this, the large-scale deep learning analysis suggested that the reliable classification of FEP based on neuroanatomical data may be around 60%. In light of these results, future studies should continue the pursuit for larger samples combined with multimodal approaches to build more reliable and informative models.

Table of contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Acknowledgements

This work was funded by a studentship awarded by the Fundação para a Ciência e Tecnologia from the Portuguese Ministry of Science, Technology and Higher Education.

I would like to thank my supervisors, especially my first supervisor, Professor Andrea Mechelli, for his insightful thoughts, kindness, guidance, and encouragement over the past years. It has been a pleasure working with you. Thank you also to Dr Stefania Tognin for her support with teaching, data analysis and my small role in PSYSCAN. Thank you to Dr Isabel Valli for getting me here in the first place.

I am also grateful to the PSYSCAN committee and the principal investigators of the studies included in this thesis for granting me access to a large amount of data; a privilege that not many PhD students have had.

Thank you to all my friends and colleagues at the IoPPN, especially the Department of Psychosis Studies for making my time here so great. Thank you to my teammates Ryan, Rafael, Lea, Cristina, Walter and Lucy for the brilliant discussions about everything, from maps to sewing and gaming. A much-needed distraction during the last few months. Thank you, Cristina, for the interesting discussions and support. Your genuine and relentless enthusiasm about research is inspiring. A special thank you to Walter, whose patience, dedication and thoughtful guidance are responsible for everything I know about machine learning.

Thank you to my dearest friends João, Cristiana, Joana, Patricia, Rita, and Sandra for their support despite the distance. Your love and support make me feel at home. I would also like to thank my brother Paulo and sister-in-law Annelies for their encouragement these last few years, and my nephew Rafael and niece Leonor for showing me how to enjoy the simple things in life. Thank you to my partner, João, who has been my main source of comfort in all my adventures. Finally, I would like to thank my parents, Celeste and Diamantino, without whom none of this would have been possible. Thank you for your unconditional support throughout all my life and for teaching me the values of hard work, dedication, and integrity.

My role in the work described

Analysis planning: I planned all the aims, hypotheses and analyses carried out as part of this thesis.

Data analysis: I carried out most of the imaging data preprocessing along with Cristina Scarpazza, a post-doctoral researcher from the same team. The scripts for the machine learning analysis were written in collaboration with Walter Pinaya, also a post-doctoral researcher from the same team. I conducted all the univariate and machine learning data analysis.

Publications: I drafted the first version of all manuscripts, implemented the suggestions made by the co-authors and revised the manuscript according to the reviewers' comments during the publication process.

List of publications derived from this thesis

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58-75.

Vieira, S., Gong, Q. Y., Pinaya, W. H., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Van Haren, N. E. … & Mechelli, A. (2019). Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophrenia bulletin*.

Vieira, S., Gong, Q. Y., Pinaya, W. H., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Van Haren, N. E. … & Mechelli, A. Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis. *Psychological Medicine.*

Vieira, S., Gong, Q. Y., Pinaya, W. H., Scarpazza, C., Crespo-Facorro, B., Van Haren, N. E. … & Mechelli, A. Using deep learning and structural data to identify first-episode psychosis: a multi-centre mega-analysis. In preparation.

Chapter 1

Background and literature review

This chapter is based on the chapters *Introduction to Machine Learning* and *Introduction to Deep Learning* published in the book Machine Learning: Methods and applications to Brain Disorders.

Vieira, S., Pinaya, W. H. L., Mechelli, A. (2020). Introduction to Machine Learning. In A. Mechelli and S. Vieira (Eds), *Machine Learning: Methods and applications to Brain Disorders*. Elsevier and Academic Press.

Vieira, S., Pinaya, W. H. L., Mechelli, A. (2020). Introduction to Deep Learning. In A. Mechelli and S. Vieira (Eds), *Machine Learning: Methods and applications to Brain Disorders*. Elsevier and Academic Press.

1.1. Introduction to psychotic disorders

Psychotic disorders are amongst the most debilitating mental disorders (C. J. L. Murray et al., 2012; Walker, McGee, & Druss, 2015). With a lifetime prevalence of approximately 3% (Perälä et al., 2007), psychotic disorders constitute one of the costliest disorders, representing about 9% of all economic costs of brain disorders in Europe (Olesen et al., 2012). The experience of a psychotic episode can involve a constellation of symptoms, typically categorized in positive (or 'reality distortion') symptoms such delusions, hallucinations and formal thought disorder, and negative (or 'psychomotor poverty') symptoms such as problems with emotion experience (e.g. anhedonia, avolition, apathy) and emotion expression (e.g. blunted/restricted affect) (van Os & Kapur, 2009).

Current diagnostic systems such as the International Statistical Classification of Diseases and Related Health Problems (ICD-10) (World Health Organization, 2004) and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013) classify the psychotic illness into a myriad of categories that describe how symptoms can be clustered to allow grouping of patients. These categories are often further grouped into non-affective psychosis, which includes schizophrenia, schizophreniform disorder, schizoaffective disorder, delusional disorder, brief psychotic disorder, substance induced psychotic disorder; and affective psychosis which includes depression/bipolar disorder with psychotic features. The classification of the observed symptoms into one of these psychotic disorders mainly depends on the number and duration of symptoms, presence or absence of affective symptoms and substance use.

Within psychotic disorders, schizophrenia emerges as one of the most disabling mental illnesses with a devastating impact on the individual (J. F. Hayes, Marston, Walters, King, & Osborn, 2017; Olfson, Gerhard, Huang, Crystal, & Stroup, 2015; Simon et al., 2018), their carers (L. Hayes, Hawthorne, Farhall, O'Hanlon, & Harvey, 2015) as well as wider society (Jin & Mosweu, 2017). With an estimated global annual incidence of 15.2 per 100,000 people (McGrath, Saha, Chant, & Welham, 2008), and lifetime prevalence of 0.40% (Saha, Chant, Welham, & McGrath, 2005), schizophrenia affects approximately 1% of the population worldwide (World Health Organization, 2008), albeit at different rates depending on geographic location and socio-demographic

background (der Werf et al., 2014). Considerable efforts have been made over the past half a century to investigate a range of neurobiological (Howes, McCutcheon, & Stone, 2015), genetic (Consortium et al., 2014), cognitive (Fett et al., 2011) and environmental (van Os, Kenis, & Rutten, 2010) factors that may lead to a better understanding of the disorder. Nevertheless, despite these efforts, the pathophysiology of schizophrenia is not yet fully understood and there is no evidence that the individual and social burden associated with the illness has subsided (J. F. Hayes et al., 2017; Hjorthøj, Stürup, McGrath, & Nordentoft, 2017).

The search for markers of schizophrenia has been mostly focused on patients suffering from this disorder for several years, i.e. chronic schizophrenia (ChSz). This it has made it difficult to disentangle which alterations observed in these patients are an intrinsic feature of the disorder or if they represent a secondary effect of factors associated with long duration of illness such as pharmacological treatment (Huhtaniska et al., 2017; Nielsen et al., 2015; A Vita, De Peri, Deste, Barlati, & Sacchetti, 2015) or chronicity (Olabi et al., 2011; A Vita, De Peri, Deste, & Sacchetti, 2012). Therefore, many researchers have recently begun focusing their efforts on those thought to be in the illness' earliest stages (McGorry, Killackey, & Yung, 2007). Within this relatively new line of research, many have focused primarily on those who have experienced a recent first episode of psychosis (FEP). This typically involves investigating individuals experiencing their first episode of schizophrenia, although it is also common to include other psychotic disorder, as defined by the ICD or DSM.

This shift has been mostly driven by the assumption that the effects associated with long duration of a psychotic illness will be reduced, if not at all absent, in individuals at this stage of illness, thus enabling a better access to the primary mechanisms underlying the illness. It is therefore ultimately hoped that this effort in investigating the onset of the illness will lead to early detection and treatment options that minimise and/or prevent the onset of established recurrent psychotic episodes.

1.2. Neuroanatomical abnormalities in first episode psychosis

The first evidence of brain structural changes in psychotic disorders dates back to the 1970's

when Johnstone et al. (1976) reported an increased ventricular volume in ChSz using computed axial tomography. Since then, a vast number of studies quickly followed in the quest to find an anatomical marker of schizophrenia. This surge was further propelled by the development of magnetic resonance imaging (MRI) which later became one of the most commonly used neuroimaging techniques, mostly due to its non-invasive nature and lack of radiation. Using this approach, most studies investigating neuroanatomical abnormalities in psychotic disorders have focused on the measurement of grey matter (GM) volume or density of cortical and subcortical brain regions between patients and controls, although a growing number of studies have also investigated cortical thickness.

Forty years since the first study, the presence of structural brain abnormalities in ChSz has been well established (Glahn et al., 2008; Haijma et al., 2013; Honea, Crow, Passingham, & Mackay, 2005; Shepherd, Matheson, Laurens, Carr, & Green, 2012; Wright et al., 2000). Reductions in GM volume, primarily in the frontal and temporal lobes, and enlargement of the lateral ventricles are among the most replicated findings (Glahn et al., 2008; Haijma et al., 2013; Wright et al., 2000). Changes in cortical thickness are less consistence, however there is evidence supporting a widespread cortical thinning across the brain, mostly in fronto-temporal regions including the fusiform, parahippocampal, inferior temporal gyri, and insula; as well as an increased thickness in mostly parietal regions (van Erp et al., 2018). Critically, these changes have been shown to be associated with chronicity (Olabi et al., 2011; van Erp et al., 2018; Vita, De Peri, Deste, & Sacchetti, 2012) and anti-psychotic medication (Radua et al., 2012; Shah et al., 2017; Antonio Vita et al., 2015). This suggests that the evidence from ChSz, although useful to describe the extent of neuroanatomical alterations observed in chronic patients, they may not reflect the changes associated with the emergence of the illness itself. Based on this premise, several studies have focused on the investigation of neuroanatomical alterations in individuals experiencing their first episode of a psychotic disorder. The following sections provide an overview of the investigation of grey matter volume and cortical thickness in FEP.

1.2.1. Grey matter volume

Most studies investigating GM volume in the early stages of psychosis have used whole-brain

voxel-based morphometry (VBM), an automated computerised technique that allows voxel-wise analysis of anatomical brain images (Ashburner & Friston, 2000). In one of first the VBM studies in FEP, Job et al. (2002) found a GM reduction in the anterior cingulate, medial frontal lobe, middle temporal gyrus, postcentral gyrus, as well as in the limbic lobe in patients relative to controls. Since then, several studies have quickly followed, and the number of neuroanatomical studies is now far greater than any other imaging modality. In the first attempt to summarise this evidence, Ellison-Wright et al. (2008) conducted the first meta-analysis of VBM studies in FEP. It was reported that, relative to controls, patients exhibited a significant reduction in GM volume in the thalamus, left uncus/amygdala region, bilateral insula and anterior cingulate. In addition, FEP patients also had a widespread pattern of GM volume increases including the putamen, insula, cerebellum and the superior frontal, middle frontal, superior temporal, precentral, lingual and fusiform gyri. Several other meta-analyses have been conducted in an effort to summarize the large number of VBM studies that followed (Table 1.1). In the largest meta-analysis yet, Radua et al. (Radua et al., 2012) analysed the findings from 25 studies and found GM reductions in the insula, middle temporal, superior temporal, precentral, medial frontal and cingulate gyri. In the latest meta-analysis, Gao et al. (2018) investigated 16 studies of anti-psychotic naïve FEP patients and found a widespread pattern of deficits including the superior and middle temporal, inferior frontal, fusiform, parahippocampal gyri, as well as in the cerebellum, hippocampus and insula.

Table 1.1. Main findings from the last meta-analyses of VBM studies in FEP.

	FEP<HC	FEP>HC
Ellison-Wright et al. (2008)	*Bilateral:* uncus/ amygdala, insula, caudate, inf frontal g; *Left:* postcentral g, cerebellum; *Right:* cingulate g, precentral g, thalamus	*Bilateral:* sup frontal g, precentral g; *Left:* putamen, mid frontal g, lingual g, cerebellum, sup temporal g; *Right:* fusiform g, insula
Chan et al. (Chan, Di, McAlonan, & Gong, 2011)	*Bilateral:* insula, sup temporal g, inf frontal g, med frontal g, postcentral g; *Left:* amygdala, mid frontal g, uncus, transverse temporal g; *Right:* precentral g, cingulate g, caudate, cerebellum	-
Bora et al. (2011)	*Right:* insula, sup temporal g, med frontal g, cingulate g	-
Fusar-Poli et al. (2012)	*Right:* sup temporal g; *Left:* insula, cerebellum	-
Radua et al. (2012)	*Bilateral:* insula, mid temporal g, sup temporal g, precentral g, med frontal g, cingulate g	*Right:* lingual g; *Left:* precentral g
Fusar-Poli et al. (2014)	*Right:* sup temporal g; *Left:* insula, cerebellum, cingulate	-
Shah et al. (2017) (AN)	*Bilateral:* insula, sup temporal g, rolandic operculum, heschl g, putamen; *Right:* mid temporal g, cingulate; *Left:* inf frontal g, postcentral g, inf frontal g, supramarginal g, fusiform g, cerebellum, parahippocampal g	*Left:* inf parietal g, paracentral g, precentral g, supplementary motor area
Shah et al. (2017) (AT)	*Bilateral:* cingulate, insula, precentral g, sup temporal g, postcentral g, inf frontal g, supramarginal g, sup frontal g; *Right:* ronaldic operculum; *Left:* mid temporal g, inf temporal g	*Right:* inf occipital g, mid occipital g, sup occipital g, sup frontal g, precentral g,
Gao et al. (2018)	*Bilateral:* insula, sup temporal g; *Right:* rolandic operculum, mid temporal g, supramarginal g; *Left:* fusiform g, cerebellum, parahippocampal g, hippocampus, inf frontal g, inf parietal g	*Right:* lingual g, cerebellum, sup frontal g

AN: anti-psychotic naïve; AT: anti-psychotic treatment; g: gyrus; inf: inferior; sup: superior; mid: middle; med: medial.

Taken collectively, these meta-analyses show that grey matter reductions in the early stages of psychosis tend to be widespread across the brain (Figure 1.1).

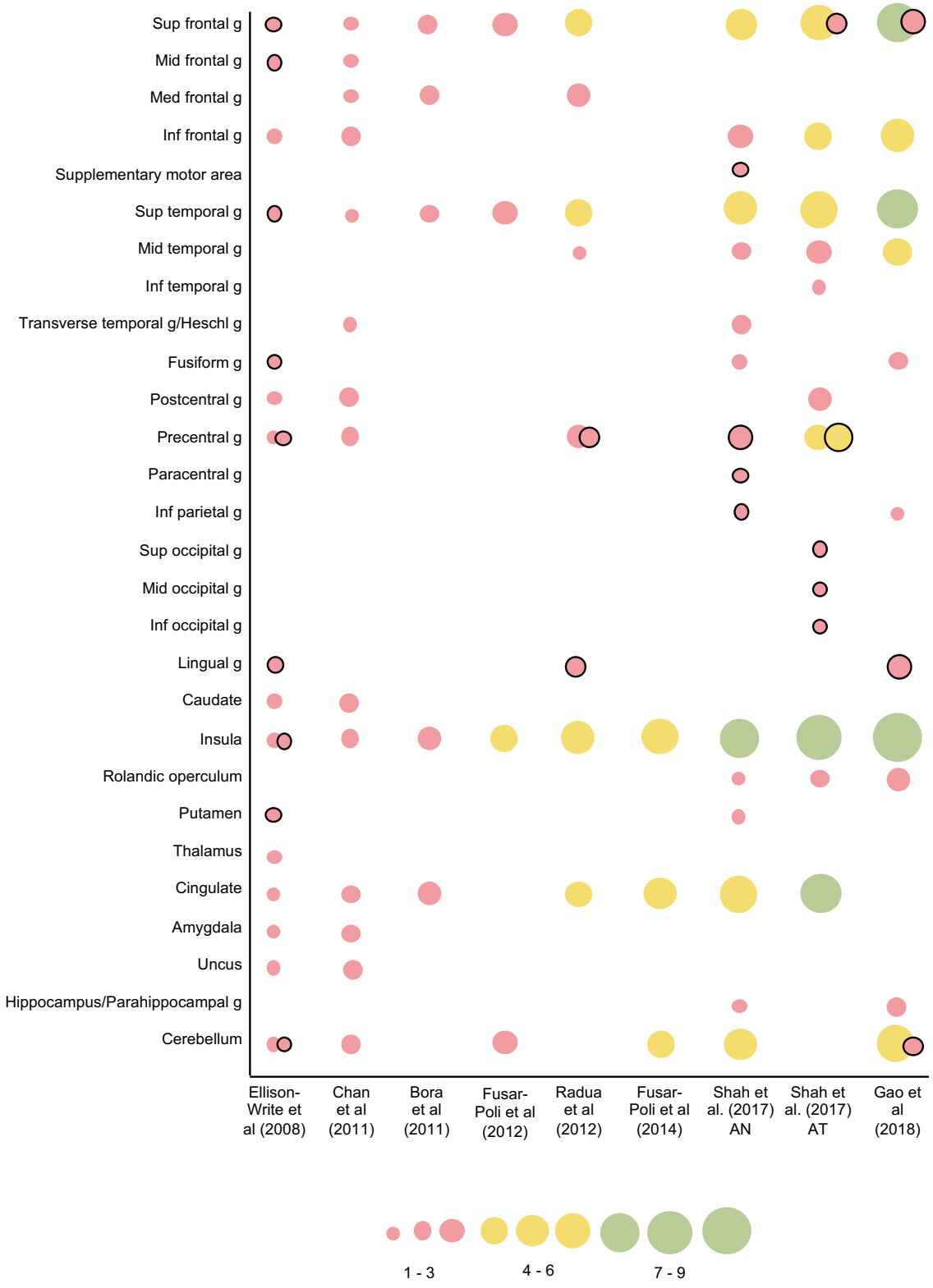


Figure 1.1. Brain regions identified in the meta-analyses in Table 1.1. The plot shows all the unique regions identified across all meta-analyses and the cumulative frequency of each region over time. Increases and

decreases of GMV are shown with and without a black circle, respectively. Note: there is substantial overlap in the studies included between meta-analyses and some of them are limited to a specific topic, for example anti-psychotic naïve (Fusar-Poli et al., 2012; Shah et al., 2017) or multimodal imaging (Radua et al., 2012). AT: anti-psychotic treatment; AN: anti-psychotic naïve.

Reductions in the superior frontal and temporal gyri, insula and cingulate are amongst the most consistent findings across meta-analyses, albeit the exact location of these reductions varies considerably. Volume reductions in several other brain regions, such as middle temporal gyrus, inferior and medial frontal gyri, pre and postcentral gyri and cerebellum have also been implicated, although less consistently. Findings of GM increases are far less consistent with some evidence towards parietal and frontal regions.

In conclusion, although fewer in number, neuroanatomical studies in FEP conducted to date suggest that the widespread alterations in GM observed in the ChSz appear to be already present in early psychosis, albeit to an less severe degree (Ellison-Wright et al., 2008; Torres et al., 2016). Notability however, there seems to be a significant heterogeneity in findings between individual studies. This is particularly salient with respect to GM deficits in the insula and fusiform gyrus (X. Gao et al., 2018; Shah et al., 2017), as well as in the middle and inferior frontal, precentral, superior and middle temporal gyri (Shah et al., 2017). Such heterogeneity may stem from methodological issues, such as the use different imaging methods, the use of small sample sizes or different recruitment criteria (Bora et al., 2011; Fusar-Poli et al., 2014), or from the neuroanatomical heterogeneity between patients (Brugger & Howes, 2017).

1.2.2. Cortical thickness

In addition to alterations in GM volume, the investigation of neuroanatomical abnormalities in psychosis can also be expressed in terms of changes in cortical thickness. While the analysis of GM volume provides a mixed measure of GM, including cortical surface area or cortical folding and cortical thickness, the analysis cortical thickness specifically targets the presence of cortical atrophy (Hutton, De Vita, Ashburner, Deichmann, & Turner, 2008; Hutton, Draganski, Ashburner, & Weiskopf, 2009). Therefore, the two approaches provide complementary information and are

recommended to use in combination to provide a more complete representation of neuroanatomical changes (Hutton et al., 2009).

Studies in FEP patients have shown consistent evidence of a widespread thinning of the cortex. Within the frontal lobe for example, Asmal et al. (2018) revealed a reduction in cortical thickness in several areas of the orbitofrontal, superior, middle frontal regions of the brain in a sample of 92 FEP and 92 controls. Similar findings were also reported by Xiao et al. (2015) and Venkatasubramanian et al. (2008) who found significant thinning of the orbitofrontal cortex and inferior frontal gyrus. Further cortical thinning has also been observed in across the temporal lobe, including in the superior, inferior and middle temporal gyri (Benetti et al., 2013; Qiu, Gan, Wang, & Sim, 2013; Scanlon et al., 2014; Song et al., 2015). Within the occipital lobe, reductions in cortical thickness have also been observed in the cuneos (Asmal et al., 2018; Qiu et al., 2013; Xiao et al., 2015) of FEP patients relative to controls, as well as in occipitotemporal regions, including the lingual (Asmal et al., 2018) and fusiform gyri (Asmal et al., 2018; Haring et al., 2016; Qiu et al., 2013). Finally, further cortical thinning has also been found in the parietal lobe in the precentral and postcentral gyri (Xiao et al., 2015) and precuneos (Asmal et al., 2018). In addition to cortical GM regions, thickness deficits have also been identified in a range of subcortical structures including the parahippocampal gyrus (Asmal et al., 2018; Qiu et al., 2013; Schultz et al., 2010), insula (Haring et al., 2016; Song et al., 2015) and anterior cingulate (Fornito et al., 2008; Haring et al., 2016). Taken together, the evidence currently available suggests the presence of multiple regions of altered thickness within both cortical and subcortical structures that could potentially be used as identifiable markers of FEP patients.

1.3. Mega-analysis of neuroanatomical data in psychiatric neuroimaging

Despite the impressive advances in the understanding of the neurobiological basis of psychiatric disorders in the last decades, there are growing concerns about the reliability and reproducibility of most findings (Anonymus, 2013). Perhaps the most noteworthy source of concern are the small sample sizes that dominate most of the neuroscientific literature (Button et al., 2013). It has been argued that such lower powered studies are more prone to false positives (Button et al., 2013) and more likely to yield heterogeneous findings (IntHout, Ioannidis, Borm, & Goeman, 2015)

compared to larger and more powerful studies. These concerns have led the neuroimaging community to acknowledge the pressing need for larger samples. However, this comes with several challenges including the limited time for recruitment impose by funding grants, financial costs, training and availability of patients in a given geographic location.

In light of such difficulties, the neuroimaging community is embracing the Big Data movement as a way to achieve sample sizes that would not be feasible within a single research site (Iniesta, Stahl, & McGuffin, 2016; Mahmoodi, Leckelt, van Zalk, Geukes, & Back, 2017; Poldrack & Gorgolewski, 2014; Van Horn & Toga, 2014). The last decade has seen a growing number of neuroimaging consortia. Notable examples include the ENIGMA (Bearden & Thompson, 2017), ADNI (Mueller et al., 2005b) and UK Biobank (Sudlow et al., 2015) which have resulted in unprecedented sample sizes in schizophrenia (van Erp et al., 2018), bipolar disorder (Hibar et al., 2018), major depressive disorder (Schmaal et al., 2017; Shen et al., 2017), autism (Postema et al., 2019) and Alzheimer's disease (Weiner et al., 2017). In what is perhaps the most prominent example in psychotic disorders, the ENIGMA consortium has led to unprecedented sample sizes in ChSz research, with two recent studies of neuroanatomical cortical abnormalities in 4474 patients and 5098 controls (van Erp et al., 2018), and subcortical changes in a smaller, albeit still impressive, sample of 2028 patients and 2540 controls (van Erp et al., 2016). These sample sizes steam for an organized effort to combine the results from several single-site studies, where each site uses the same pipeline for data preprocessing and analysis; once analysed, single-site results are pooled and summarized in a meta-analysis. Data-sharing initiatives are also increasing rapidly, with over 40 online repositories for neuroscientific data in 2015 (Eickhoff, Nichols, Van Horn, & Turner, 2016; Ferguson, Nielson, Cragin, Bandrowski, & Martone, 2014). In the first effort to combine several publicly available datasets of ChSz, Gupta et al. (C. N. Gupta et al., 2015) analysed 784 patients and 936 healthy controls collected from 23 sites. More recently, Rozycki (2018) analysed data from 5 sites totalling 448 healthy controls and 387 ChSz patients.

Taken collectively, these studies represent the initial steps of a movement that will hopefully pave way for more reliable and reproducible findings in psychiatric neuroimaging in general, and psychosis in particular. Although mega-analyses raise important challenges, such as the

integration of data from different scanners or greater heterogeneity amongst participants, larger samples are more likely to be more representative of the illness and thus carry more translational potential. Critically, similar mega-analytic efforts focussed on the initial stages of psychosis, when the effects of confounders are minimal, are still non-existent and, as described in the previous section, evidence is still reliant on small to modest sized studies.

1.4. Machine learning

The emergence of neuroimaging in the 1990s has led to impressive advances in the understanding of brain disorders including both psychiatric and neurological disease. However, the traditional case-study design that has dominated most of the neuroimaging literature for the past three decades, rooted on lesion studies (Scoville & Milner, 1957) and theories of modularity (Fodor, 1983), was designed to test hypotheses about neural mechanisms and without translational goals in mind. Therefore, while much progress has been made, very few results have been incorporated into clinical practice (Dazzan, 2014; Prata, Mechelli, & Kapur, 2014; Woo, Chang, Lindquist, & Wager, 2017). In an attempt to bridge this gap between research and clinical practice, the neuroimaging community has developed a growing interest in machine learning in the hope that this approach will circumvent some of the limitations of classical statistics that are hindering the translational application of findings. There are at least four ways in which machine learning breaks with classical statistics that may help achieve this: 1) it allows individual rather than group-level inferences, 2) it is inherently a multivariate, as opposed to univariate, approach, 3) it focuses on prediction and generalizability, instead on explained variability, and finally 4) it is more sensitive to heterogeneity in the data, rather than creating a 'typical' average participant. This section expands on each one of these differences, after proving a brief definition of machine learning.

1.4.1. Definition

Machine learning is an area of artificial intelligence that has emerged as part of the ongoing quest for building intelligent machines that are capable of learning. Although the term 'machine learning' was coined in 1959 (Samuel, 1959), machine learning only emerged as an area of artificial intelligence in the 1980s (Langley, 2011, 2016). Machine learning relies on the intersection of

several disciplines including computer science, engineering, mathematics, statistics, psychology, and neuroscience. Perhaps as a result of its short history and interdisciplinary nature there has been much debate about the definition of machine learning. Nevertheless, machine learning is usually referred to as an area of artificial intelligence that is concerned with identifying patterns from data and use the same patterns to make predictions about unseen data (Mitchell, 1997). From here it follows that the main outcome of machine learning is a measure (proxy) of generalizability: the extent to which a model is capable of outputting correct predictions when presented with new data, based on learned rules from previous exposure to similar (but not the same) data (Domingos, 2012).

1.4.2. Machine learning versus classical statistics

1.4.2.1. Individual-level versus group-level inferences

One of the main reasons for the existing gap between research and clinical practice is that the former has been dominated by methods that only allow inferences at group-level (e.g. a group of psychosis patients have larger ventricles than a group of controls); whilst a clinician has to make diagnostic or treatment decisions at the level of the individual. A key reason why machine learning is gaining considerable attention amongst the research and medical communities is that it promises to bridge this gap. By learning patterns in the data that best distinguish between patients with a certain disease of interest and healthy individuals, for example, it is possible to estimate the likelihood that a new set of data acquired from an individual belongs to a patient or a healthy individual. Similarly, by learning patterns in the data that best distinguish between patients who benefit from a certain treatment and patients who do not benefit from it, it is possible to estimate the likelihood that a new set of data acquired from an individual belongs to a “responder” or a “non-responder”. Therefore, machine learning opens new possibilities in personalized medicine, by allowing the development of novel tools that could be used to inform diagnostic and treatment decision-making in everyday clinical practice.

1.4.2.2. Multivariate versus univariate analysis

The vast majority of clinical neuroimaging studies are based on mass-univariate methods, i.e. a separate statistical test is performed to investigate each variable of interest. In VBM studies,

statistical parametric mapping is typically used to perform a large number of voxel-wise comparisons between groups, without considering possible interaction between voxels. However, this approach is not in line with the current understanding of brain anatomy and function (Biswal et al., 2010; Fox et al., 2005). Indeed, as explained in the previous section, neuroanatomical abnormalities in psychosis are characterized by subtle and widespread changes, rather than isolated focal alternations. Machine learning, on the other hand, is inherently a multivariate approach; it is capable of taking the relationship between multiple variables inputted into the same model into account, thereby allowing greater sensitivity to subtle and widespread changes in brain anatomy.

1.4.2.3. Prediction and generalizability versus explained variability

Classical inferential statistics is mainly concerned with elucidating the relationship between observed phenomena of interest, for example to what extent changes in the brain anatomy explain severity of symptoms (Yarkoni & Westfall, 2017). In neuroimaging, for instance, studies typically involve the use of the general linear model, which estimates the strength of the association between independent and dependent variables, and returns a measure of explained variability or goodness of fit, i.e. the extent to which a statistical model accounts for the variation in the data (Friston et al., 1994). It is often assumed that models with high explanatory power or goodness of fit have high predictive power when applied to real-world cases. However, from a statistical perspective, the model that best describes a set of observations at group-level will not necessarily be the most successful at predicting real-world outcomes (Arbabshirani, Plis, Sui, & Calhoun, 2017; Shmueli, 2010). This is because, whilst a statistical model may achieve a high explanatory power or goodness of fit when fitted to a particular dataset, it will likely incorporate the unique characteristics of the dataset (Yarkoni & Westfall, 2017). When a dataset is large enough to ensure representativeness of the population from which it was drawn, this is less likely to be an issue. However, in the vast majority of clinical neuroimaging research which involves studies with small samples, representativeness is not guaranteed, and models will likely capture fluctuations in the data that are unique to a particular sample; this is known as ‘overfitting’. The extent to which findings from a single study are generalizable to other samples is not usually addressed in studies using classic inferential statistics (Bzdok & Yeo, 2017). Importantly, machine learning does not

necessarily solve the issue of generalizability. However, it does at least attempt to measure it. In fact, building models capable of accurate predictions in unseen data is the fundamental goal of machine learning.

1.4.2.4. Heterogeneity versus 'typical patient'

It is well known that psychiatric and neurological disorders tend to be heterogenous in terms of underlying neuroanatomical and neurofunctional alterations, clinical presentation and progression over time (Holmes & Patrick, 2018; Insel et al., 2010; Wardenaar & de Jonge, 2013). However, most advances in brain disorders research, including rigorous clinical trials for example, are based on the idea of a 'typical patient', which masks individual variability. In contrast, by looking for a multivariate pattern across a group of individuals during training, machine learning is sensitive to heterogeneity in the data. However, integrating heterogeneity in individual-level modelling can be challenging, as it becomes more difficult to find patterns that are relevant to the task at hand above and beyond individual heterogeneity (Schnack, 2017). In light of the current trend to recruit larger and larger sample sizes, data is likely to become more heterogeneous. This is in sharp contrast with the traditional case-control approach where, ideally, the patient group and the control group are expected to be as homogeneous as possible. On the other hand, larger samples are likely to be more representative of the illness and thus carry more translational potential in real-word clinical practice.

1.4.2.5. Data-driven versus hypotheses-driven models

Historically, research into brain disorders has been heavily based on deductive (top-down) or theory-driven approaches, where carefully thought-out and well-defined hypotheses are tested and ultimately confirmed or rejected. Having a priori hypothesis is considered paramount as it minimises the risk of false positive findings and post-hoc explanations (Kitchin, 2014). More recently, increasing access to large datasets combined with technological advances have propelled the emergence of data-driven approaches, such as machine learning, where insights are generated purely from data in a bottom-up fashion. Contrary to classical statistics, where the aim is to test a priori hypotheses whilst making significant assumptions about the data (e.g. linearity, normal distribution), in machine learning the main premise is to 'let the data speak for

itself, whilst making as few assumptions as possible about the data (Bzdok, 2017; Jordan & Mitchell, 2015; Mahmoodi et al., 2017).

1.4.3. Bias-variance trade-off, model assumptions and regularisation

There are multiple ways in which a model can learn patterns from the data, resulting in a multitude of machine learning algorithms. A common taxonomy organizes the different approaches according to the style of learning. Based on this categorization, machine learning methods can be grouped into four different types of learning: supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning is by far the most commonly used approach in general and in psychiatric neuroimaging, and also the type of learning used in this thesis. The main aim of any supervised machine learning model is to build a function that maps the observed data (i.e. features) and a target variable capable of generalizing beyond the set of data used to develop this function. Essential to the implementation and interpretation of any supervised machine learning model are the concepts of bias-variance trade-off, regularisation and model assumptions. This section briefly introduces these concepts.

Building a successful machine learning model is often a continuous process, where increasingly more complex models (e.g., models with large number of parameters to estimate) are developed to achieve better performances. As complexity increases, however, there are two main sources of error - bias and variance - that need to be balanced (Figure 1.2). Bias arises when the model learns a faulty assumption in the data. In the presence of high bias, the algorithm will not be able to model the relationship between features and target correctly; this is known as underfitting. As shown in Figure 1.2A, although data tends to plateau, the model assumes there is a linear association between the variables. A model that uses inefficient or uninformative features, a very small number of observations, or a too simple algorithm, for example, is likely to be too simple to capture any meaningful patterns and result in a highly biased model. On the other hand, variance results from modelling detailed fluctuations in the data. An algorithm with high variance will capture specific aspects of the training data that do not generalize well in the test set; this is known as overfitting. This happens when, for example, there are too many features relative to the number of observations or when very complex models are implemented (Figure 1.2A). The way

in which model complexity affect bias and variance is known as the bias-variance trade-off (Figure 1.2B).

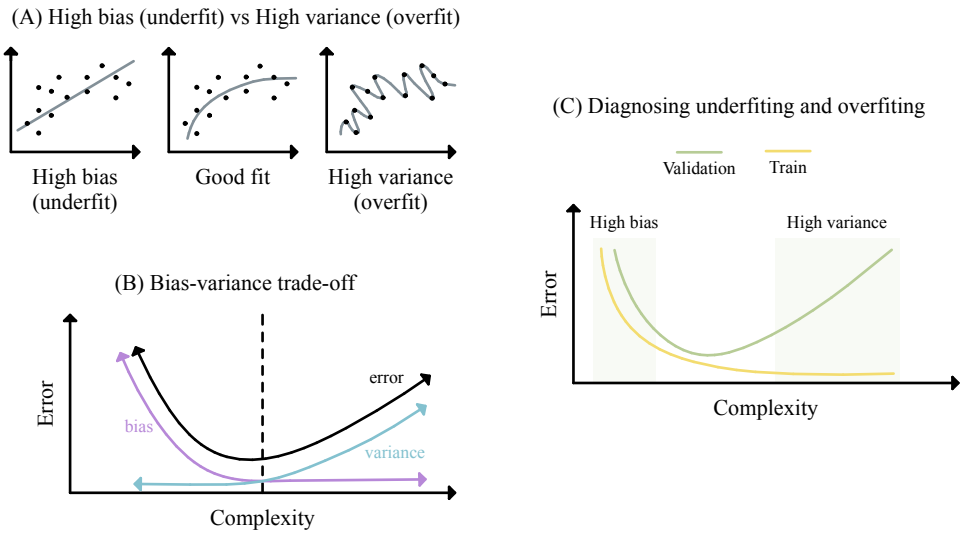


Figure 1.2. Bias-variance trade-off. (A) Three models with different levels of bias and variance. The model in the left has low variance but high bias, while the one on the right has high variance but low bias. An optimal solution would be a model with a good balance between bias and variance. (B) Bias-variance trade-off. As model complexity increases, variance increases and bias decreases. Ideally, bias and variance can be balanced, and the algorithm will achieve convergence (when additional training will not improve the model). (C) Diagnosing underfitting and overfitting. By assessing the error in the training and validation sets, it is possible to establish whether a model is under- or overfitting the data.

When an algorithm performs poorly, this will most likely be due to either a high bias or high variance issue. Therefore, to improve the model's performance, it is necessary first to identify whether a model is underfitting or overfitting the data (Figure 1.2C). When the model is too simple, error in both training (data used to develop the model) and validation (data used to test the model) sets is high, indicating that the model is not a good fit, i.e., the model is underfitting the data. On the other hand, when the model is too complex, it fits the training data very well, i.e., the error is close to zero; however, the error in the validation set is high, indicating overfitting and poor generalizability. Several reasons may help explain poor performance, either due to under- or overfitting. One important reason may be that the data does not meet the assumptions that underly the inner workings of the model. A simple example here is the common assumption among several traditional supervised machine learning algorithms that the target variable can be

predicted with a linear combination of the features. This means that if this assumption is wrong, the model will underestimate the strength of the relationship between the two. Different machine learning models have their own specific assumptions and if the data does not meet these assumptions, this will likely lead to poor performance. Overfitting is another common reason for poor performance. Since the chances of overfitting increase with model complexity, a common approach to help minimize (or even prevent) overfitting involves penalizing model complexity. This is referred to as regularization; a group of strategies that force a model to favour simpler (i.e. less complex) solutions. The use of regularization is common in psychiatric neuroimaging research. The dimensionality of the image data is often much larger than the number of observations. In a typical neuroimaging study, there are potentially hundreds of thousands, or even millions, of dimensions (e.g. voxels), whereas the number of observations is typically of the order of dozens to hundreds, implying that machine learning built with neuroimaging data are extremely ill-posed (i.e., have more than one solution) (Lautrup et al., 1995). Common examples of regularization techniques include ridge regression or L2, lasso or L1 and elastic net. Briefly, these strategies involve the use of weight decays to penalize models with very high weights. By forcing weights to remain low, the model becomes less dependent on the training data (i.e., performance does not rely heavily on a particular set of weights) and can better generalize to unseen data (Nowlan & Hinton, 1992). Other types of strategies also exist. For example, dropout consists of temporarily removing a random number of neurons and their respective incoming and outgoing connections from the network during training of deep learning models (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014); early stopping involves stopping training when the error in the training set stops decreasing, especially if this is accompanied by an increase in the error in the validation set, a strong indicator of overfitting (Prechelt, 1998).

1.4.4. Machine learning studies of first episode psychosis

Over the last decade several different machine learning approaches have been applied in brain disorders (Woo et al., 2017). Amongst the most popular ones are logistic regression and support vector machine, for example. The simplicity, interpretability and ease of use of these the approaches have resulted in a wealth of evidence across psychiatric neuroimaging in search of initial findings that could ultimately led to translational tools (Orrù, Pettersson-Yeo, Marquand,

Sartori, & Mechelli, 2012; Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015; Woo et al., 2017). In psychosis, several studies have been able to successfully distinguish between ChSz patients and healthy individuals based on neuroanatomical data (Kambeitz et al., 2015; Zarogianni, Moorhead, & Lawrie, 2013). However, it is unclear to what extent the distinction between two groups could be influenced by structural abnormalities associated with typical confounding variables in ChSz, such as long duration of illness and antipsychotic medication. Pattern classification applied to FEP could thus allow a clearer insight into the underlying mechanisms of the illness. The assumption here is that, whilst separating FEP from controls may be more difficult due to the subtler changes, the neuroanatomical differences driving the distinction between the two are likely to be more reflective of the underlying mechanisms of psychosis. Consistent with this, both disease-stage and antipsychotic medication were identified as significant moderators in a recent meta-analysis of machine learning studies in psychosis (Kambeitz et al., 2015).

Compared to ChSz, evidence from FEP studies conducted so far has been less consistent. In one of the first studies, Sun et al (2009) was able to distinguish between patients and controls with an accuracy of 86% in a sample of 36 FEP and 36 controls. To mitigate the effect of sex as a possible confounder, Takayanagi et al. (Takayanagi et al., 2010, 2011) ran separate classifiers for males and females with accuracies between 76% and 87% for males and between 81% and 83% for females. Later, Borgwardt et al. (2013) were able to classify patients and controls with an impressive accuracy of 87%. Shortly after however, Petterson-Yeo et al. (2013) reported a much lower result of 63%. More recently, Xiao et al. (2017) successfully distinguished FEP and controls in one of the largest studies yet, with 163 anti-psychotic naïve patients and 163 controls, based on measures of thickness and surface area of several cortical brain regions with 82% and 85% accuracy, respectively. Meanwhile, Winterburn et al. (2017) tested several different classifiers on three popular neuroanatomical features – two measures of voxel-wise GM volume and cortical thickness – in a sample of 50 FEP and 50 controls. Most accuracies fell between 55% and 70%, and the best performance was achieved with cortical thickness with 74% accuracy. In conclusion, the evidence from machine learning studies applied to neuroanatomical data to identify the initial stages of psychosis, where the effect of confounders is minimal, is still scarce and has been

inconclusive so far.

1.5. Deep Learning

The ease of use of several machine learning methods such as SVM have propelled a vast amount of evidence during the last decade across the field of clinical neuroimaging (Woo et al., 2017). Common to all these conventional methods, however, is their limitation in processing data in its raw form. Therefore, since the performance of any machine learning method is heavily reliant on the choice of features, much of the effort that goes into developing a successful conventional machine learning pipeline is spent on carefully creating useful features from the raw data, i.e. feature engineering (Domingos, 2012). Representation learning is a class of machine learning methods that addresses this limitation by discovering the optimal set of features, i.e. representation, from the data automatically. Deep learning is a family of representation learning methods loosely inspired on biological neurons that learn increasingly abstract levels of representations obtained from combining multiple layers of interconnected nonlinear processing units known as 'artificial neurons' (Bengio, Goodfellow, & Courville, 2015). This structure results in a great flexibility that can be leveraged to create a vast number of different architectures, many of which tailored for specific purposes. Perhaps the simplest model is the general-purpose deep neural network (also known as multilayer perceptron, fully-connected neural network, or similar variations). Other popular architectures include for example, autoencoders (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010) which are typically used for dimensionality reduction, convolutional neural networks (LeCun, Bottou, Bengio, & Haffner, 1998) which are mainly used to process images and recurrent neural networks which are used to process sequential data such as speech, video or even functional imaging data.

Deep learning has seen a dramatic surge in interest during past decade in the wider research community and industry (LeCun, Bengio, & Hinton, 2015). This has been largely driven by increases in computational power and the availability of massive new datasets, which ultimately led to record-breaking performances in visual and speech recognition tasks (Graves, Mohamed, & Hinton, 2013; Krizhevsky, Sutskever, & Hinton, 2012; Le, 2013). In medicine, deep learning is also gaining considerable attention with promising results (Esteva et al., 2019; Wang, Casalino,

& Khullar, 2019), including the detection of diabetic retinopathy (Gulshan et al., 2016) and skin cancer (Esteva et al., 2017) from retinal fundus and skin images, respectively. Despite the recent outpouring of interest however, the origins of deep learning can be traced back to 1940s with the 'perceptron', one the first attempts to model the biological neuron (McCulloch & Pitts, 1943). After a long and controversial history (Schmidhuber, 2015), including the inability to solve nonlinear problems (i.e. 'XOR-problem') and the 'vanishing or exploring gradients' problem, the perceptron evolved to become a network comprised of sever 'hidden layers' connected by weights which were optimized via backpropagation, known as artificial neural networks (Durstewitz, Koppe, & Meyer-Lindenberg, 2019). However, such networks were able to handle only a limited number of layers. It was only in the 2000s that researchers developed a new approach for training artificial neural networks that allowed the inclusion of several hidden layers by first pre-training the network layer by layer followed by the finetuning of the entire network, resulting greater levels of complexity (Hinton, Osindero, & Teh, 2006). This breakthrough led to the development of a new family of machine learning methods known as deep learning.

In light of its ability to learn latent and abstract patterns, it has been suggested that deep learning may be of particular value to uncover complex effects within a certain modality, for example subtle, widespread and heterogenous reductions in grey matter volume; or even capture cross-modality relations, such as the interaction between genetics and neuroimaging (Calhoun & Sui, 2016; Plis et al., 2014; Schnack, 2017). In addition, given that studies have traditionally relied on mass-univariate techniques, we often lack strong hypothesis about how GM volume across the brain relate to each other or how different modalities may be related, and therefore data driven methods such as deep learning may be particularly useful (Durstewitz et al., 2019). Plis et al. (2014) is often referred to as the first study to apply deep learning in the context of clinical neuroimaging. By applying a deep belief network to structural data, authors were able to successfully distinguish controls from schizophrenia patients as well as predict the severity of symptoms in patients diagnosed with Huntington disease. This study was quickly followed by several others using a variety of different modalities and types of networks that aimed to distinguish controls from Alzheimer's disease or mild cognitive impairment (Hu, Ju, Shen, Zhou, & Li, 2016; Liu et al., 2015; Suk, Lee, & Shen, 2014), autism spectrum disorders (Hazlett et al., 2017; Heinsfeld, Franco,

Craddock, Buchweitz, & Meneguzzi, 2018) and ADHD (Kuang & He, 2014; Zou, Zheng, & McKeown, 2017); or to identify individuals suffering from mild cognitive impairment would go on to develop Alzheimer's disease (Liu, Liu, Cai, Che, et al., 2015; Suk & Shen, 2013).

A limited number of studies have also been conducted in ChSz. In addition to Plis et al (2014), Kim et al. (2016) applied a deep neural network to functional MRI data in a sample of 50 ChSz and 50 controls and was able to classify the two groups with 86% accuracy. A similar result was also found by Yan et al. (2017), in which a variation of a deep neural network was also used to distinguish between patients and controls based on functional imaging, albeit in a much larger sample of 1100 participants. In an attempt to combine structural and functional imaging data, Ulloa et al. (Ulloa, Plis, & Calhoun, 2018) built a model also based in a deep learning network capable of classifying patients and controls with an accuracy of 85% in a sample of 304 participants. In a large multi-centre study, Zeng et al. (2018) applied an autoencoder-based model to functional imaging in a sample of 734 participants. Accuracies of 85% and 81% were obtained in the multi-site pooling classification and leave-site-out classification, respectively.

In conclusion, deep learning is a promising approach capable of capturing intricate relations from the data that may be useful to detect biomarkers for psychosis. Its application to clinical neuroimaging in ChSz is still at the very early stages and there have been no studies applying deep learning to the early stages of the illness.

1.6. Aim and hypothesis

In summary, although there is already a considerable amount of studies investigating focal neuroanatomical abnormalities in FEP, the vast majority of findings come from small local studies, which may help explaining the current heterogenous evidence in the literature. This is in line with growing concerns across the wider neuroscientific community regarding the failure of replication and reproducibility of findings and subsequent calls for greater collaboration to build larger and more robust studies. Coinciding with this furthermore, are the increasing calls for translatable findings into clinical practice. As a result, there has been a growing number of studies applying machine learning to individuals at the early stages of psychosis in an attempt to build tools that

can be used to assist with decision-making in the clinical practice. Despite the recent advances in the last few years however, evidence from the yet small number of studies in FEP has been inconclusive. Finally, within this movement, deep learning has recently emerged as a promising avenue for the search of biomarkers in psychiatric neuroimaging. Although initial evidence is encouraging, more research is needed, especially in the early stages of psychotic disorders, when diagnosis may be uncertain, and treatment is yet to be decided.

In the context outlined above, the primary research questions of the current thesis and respective hypotheses were as follows:

1.6.1. Are there neuroanatomical brain differences between FEP and controls consistent across several independent sites?

My first objective was to use standard univariate analyse to examine GM volume changes in FEP relative to controls that are expressed consistently across several independent samples using a multi-centre mega-analytic approach. The following hypothesis were considered:

H1. The FEP group would show GM volume reductions in a distributed bilateral network including fronto-temporal, insular and cingulate areas compared to the control group.

H2. GM volume in the FEP group would be negatively correlated with severity of symptoms.

H3. GM volume in the FEP group would be negatively correlated with duration of illness.

H4. GM volume in the FEP group would not be correlated with anti-psychotic medication.

1.6.2. What is the evidence for deep learning applications in psychiatric and neurologic neuroimaging?

My second aim was to review the literature with respect to the applications of deep learning to neuroimaging data in psychiatric and neurologic disorders. Machine learning is relatively new to the neuroimaging community and deep learning, although a sophisticated version of the long

standing artificial neural networks, is particularly novel to the community. Therefore, in addition to providing an overview of the main deep learning architectures, I aimed to carry out a systematic survey of the current literature in terms of diagnostic and longitudinal outcome studies that used some form of deep learning, as well as highlighting its limitations and main future directions.

1.6.3. Can brain neuroanatomy be used to classify FEP and controls consistently across several independent datasets?

My third aim was to elucidate the extent to which the application of popular traditional machine learning techniques to neuroanatomical data allows distinction between FEP and controls at the individual level by putting place a series of precautions to minimise the risk of overfitting. To assess the reproducibility of the findings, the same pipelines were applied to five independent datasets. It was hypothesized that:

H5. FEP and HC would be classified with statistically significant performances ranging between 70% and 80%.

H6. Performances would remain stable across the five datasets.

1.6.4. Can deep learning be used to classify FEP and controls based on brain neuroanatomical information?

My fourth aim was to examine whether deep neural networks could classify FEP and controls based on neuroanatomical data. It was hypothesized that:

H7. Deep neural networks would be able to classify FEP and controls with statistically significant performances ranging between 70% and 80%.

H8. Deep neural networks would show a superior performance compared to traditional shallow approaches.

1.6.5. Can deep learning be used to classify FEP and controls based on brain neuroanatomical information in a large-scale analysis?

The final aim of this doctoral work was to investigate whether it would be possible to classify FEP and controls using deep neural networks in a multi-centre mega-analytic approach.

H9. Deep neural networks would be able to classify FEP and controls with statistically

significant performances around 70%.

H10. Deep neural networks would show a superior performance compared to traditional shallow approaches.

H11. The main regions driving classification would include fronto-temporal regions as well as the insula and cingulate.

1.7. Structure of the present thesis

	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
	Overview of the main methodology used in this thesis	Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis	Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications	Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence	Using deep learning and structural data to identify first-episode psychosis: a multi-centre mega-analysis	General discussion
Research questions		1.6.1	1.6.2	1.6.3 and 1.6.4	1.6.5	
Hypotheses		H1 – H4		H5 – H8	H9 – H11	

Chapter 2

Methodology

2.1. Participants

2.1.1. Study sample

A total of 1249 participants were collected from five previously published independent studies:

- Site 1: Chengdu, China (Gong et al., 2015)
- Site 2: London, England (GAP[1] study; Di Forti et al., 2009)
- Sites 3 and 4: Santander A and B, Spain (PAFIP[2] study; Pelayo-Terán et al., 2008)
- Site 5: Utrecht, The Netherlands (GROUP[3] study; Korver, Quee, Boos, Simons, & de Haan, 2012)

These datasets are also part of a larger pool of legacy data for the project PSYSCAN - Translating Neuroimaging Findings from Research into Clinical Practice, an EU-funded multi-centre study that aims to develop neuroimaging-based tool to help physicians in the management of patients with psychotic disorders[4]. Permission to use this data was obtain from the PSYSCAN committee. The final sample sizes for each study included this thesis are shown in Figure 2.1. The demographic and clinical characteristics are provided in the respective chapters.

2.1.2. Participants

Site 1: Chengdu University, China

A total of 167 patients aged 18-44 years were recruited from the West China Hospital of Sichuan University in Chengdu between 2009 and 2012. Diagnosis for first episode of schizophrenia within the previous 24 months was determined by the consensus of two clinical psychiatrists using the Structured Interview for the DSM-IV Axis I Disorder (SCID-I) (First & Gibbon, 2004). At the time of scanning, all patients were medication-naïve. A total of 163 healthy controls were recruited by poster advertisement and screened using the SCID-I to confirm the lifetime absence of psychiatric disorders, as well as interviewed and subsequently excluded if they had any known history of psychiatric illness in first-degree relatives. Participants were excluded if they met any of the following criteria: i) history of drug or alcohol abuse, ii) pregnancy, and iii) any physical illness

[1] Genetics and Psychosis
[2] Clinical Program on First-Episode Psychosis of Cantabria
[3] Genetic Risk and Outcome of Psychosis
[4] http://psyscan.eu

such as hepatitis, cardiovascular disease, or neurological disorder, as assessed by interview and review of medical records. The study was approved by the local research ethics committee and all participants provided written informed consent.

Site 2: King's College London, England

Ninety-four patients aged 18-65 years were recruited from the South London and Maudsley Foundation Trust and scanned at the Institute of Psychiatry, Psychology and Neuroscience in London between December 2005 and October 2008. All patients meeting ICD–10 criteria for a diagnosis of psychosis (codes F20–F29 and F30–F33) (World Health Organization, 2004) were invited to participate in the study; patients with a diagnosis of organic psychosis (i.e. psychosis caused by a known physical illness such as toxic-metabolic encephalopathies and stroke) were later excluded. Clinical diagnosis was established by administering the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1990). A total of 110 healthy controls were recruited through local advertisement from the same geographical areas as patients. The Psychosis Screening Questionnaire (Bebbington & Nayani, 1995) was used to exclude the presence of psychotic symptomatology or a history of psychotic illness. Participants were excluded if they met any of the following criteria: i) learning disabilities (IQ < 70 derived from the Wechsler Adult Intelligence Scale–Third Edition (WAIS III); Wechsler (1997), ii) current or past neurological illness, iii) brain injury with loss of consciousness for more than 1 hour and iv) suspected or confirmed pregnancy. Ethical permission was obtained from the Trust and the Institute of Psychiatry, Psychology and Neuroscience research ethics committee and all participants provided written informed consent.

Site 3 and 4: Santander University, Spain

Data from two Spanish sites - Santander A (N patients = 144, N healthy controls = 113) and Santander B (N patients = 145, N healthy controls = 78) - was acquired as part of the same large prospective longitudinal study on first episode psychosis in the region of Cantabria, although with two different scanners. Patients aged 15-55 years were recruited from both inpatient units and community services throughout the entire region between February 2001 and February 2005. Diagnosis of a first episode of non-affective psychosis (schizophrenia, schizophreniform disorder,

schizoaffective disorder, brief reactive psychosis, or not otherwise specified psychosis) according to DSM-IV criteria was confirmed by administering the structured interview SCID–I (First & Gibbon, 2004). Patients were recruited if there was no evidence of prior treatment with antipsychotic medication or, if previously treated, a total lifetime of adequate antipsychotic treatment of less than 6 weeks. Patients with DSM-IV based diagnoses of a psychotic disorder directly caused by a general medical condition or use of substances, mental retardation or substance dependence (except nicotine dependence) were excluded. Age and sex matched healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness, as determined by using an abbreviated version of the Comprehensive Assessment of Symptoms and History (CASH) (Andreasen, Flaum, & Arndt, 1992). Clinical records and family interview also confirmed the absence of psychosis in first-degree relatives. Ethical permission was obtained from the local institutional review board and all participants provided written informed consent.

Site 5: Utrecht University, The Netherlands

A total of 105 patients aged 16-50 years were recruited from inpatient and outpatient regional psychosis departments or academic centres in Utrecht. Diagnosis of a first episode of non-affective psychosis (schizophrenia, schizophreniform disorder, schizoaffective disorder, brief reactive psychosis, or not otherwise specified psychosis) according to DSM-IV criteria was established by administering the Comprehensive Assessment of Symptoms and History (Andreasen et al., 1992). A total of 120 healthy controls were recruited through a system of random mailings to addresses in the catchment areas of the cases and were screened for current or past psychotic disorder and first-degree family member with a lifetime psychotic disorder. The study protocol was approved centrally by the Ethical Review Board of the University Medical Centre Utrecht and all participants provided written informed consent.

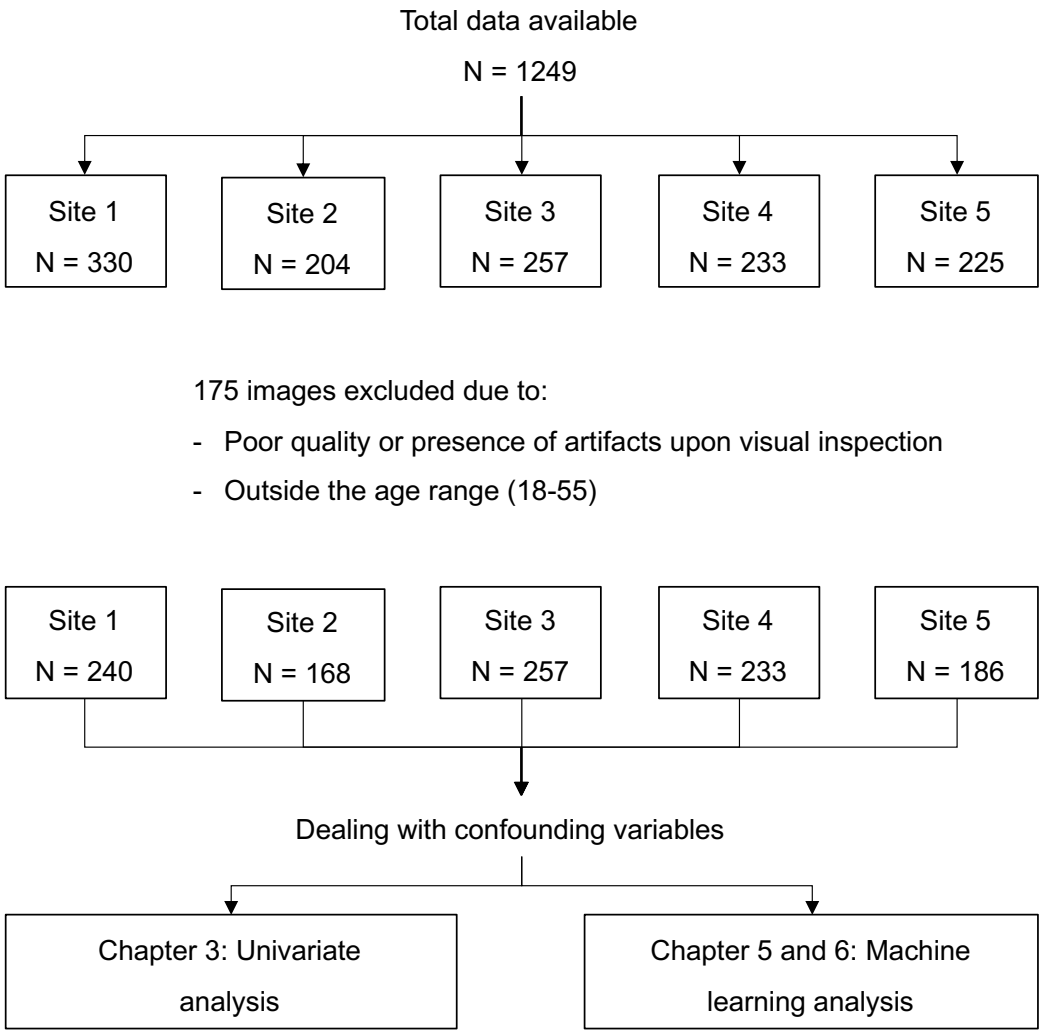


Figure 2.1. Diagram showing initial and final sample size for the univariate and machine learning analysis.

2.2. Structural magnetic resonance imaging

Magnetic resonance imaging (MRI) is a non-invasive technique that can be used to visualise different tissues of the human body *in vivo* by leveraging on the magnetic properties of the protons

within the nuclei of hydrogen atoms to produce images.

Every cell of the body contains hydrogen particles. The atom of each hydrogen particle comprises of a single charged proton which rotates, or 'spins', on its own axis with a specific direction and intensity referred to as angular momentum (Figure 2.2A). In the absence of an externally applied magnetic field, the collective magnetic moment of all spins has random orientations (Figure 2.2B). As a result, there is no overall magnetic field, i.e. the net magnetisation is equal to zero. However, when subjected to an external magnetic field (B_0) – the primary magnetic field – the spins' magnetic moments will align with this external field in one of two orientations with respect to B_0, parallel or anti-parallel (Figure 2.2C). Protons with a parallel alignment are more stable and thus possess low energy; conversely, the protons with an anti-parallel alignment are less stable and carry more energy. In this magnetisation state, there are more spins in the low-energy parallel state compared to high-energy anti-parallel state. Summing the contributions of all the spins' magnetic vectors will therefore result in a net magnetic vector (M) aligned with the longitudinal z-axis of B_0, referred to as longitudinal magnetisation (Figure 2.2D). In addition, protons will also spin along the longitudinal (z) axis of B_0 at a frequency known as the Larmor frequency; this is known as precession (Figure 2.2E). When protons precess together, this is known as in-phase, whereas when protons precess separately, this is known as out of phase.

By applying a radio frequency (RF) pulse with the same frequency at which the protons are precessing (process called resonance), a second external magnetic field (B_1) perpendicular to the z-axis is generated, in the x-y plane. This disturbs the proton alignment by forcing the protons in parallel alignment to 'flip' to the higher energy anti-parallel state, decreasing longitudinal magnetisation. In addition, it will also force all protons to precess in-phase within the x-y plane. As a result, the net magnetisation M tilts from the z-axis direction into the transverse x-y plane; this is known as transverse magnetisation (Figure 2.2F). The resulting magnetic vector M_{x-y} induces an electrical current detected by a receiver coil forming the magnetic resonance (MR) signal. The time taken between the RF pulse being applied and an MR signal being received is known as the echo time (TE) whereas the time between the application of each RF pulse is known at the repetition time (TR).

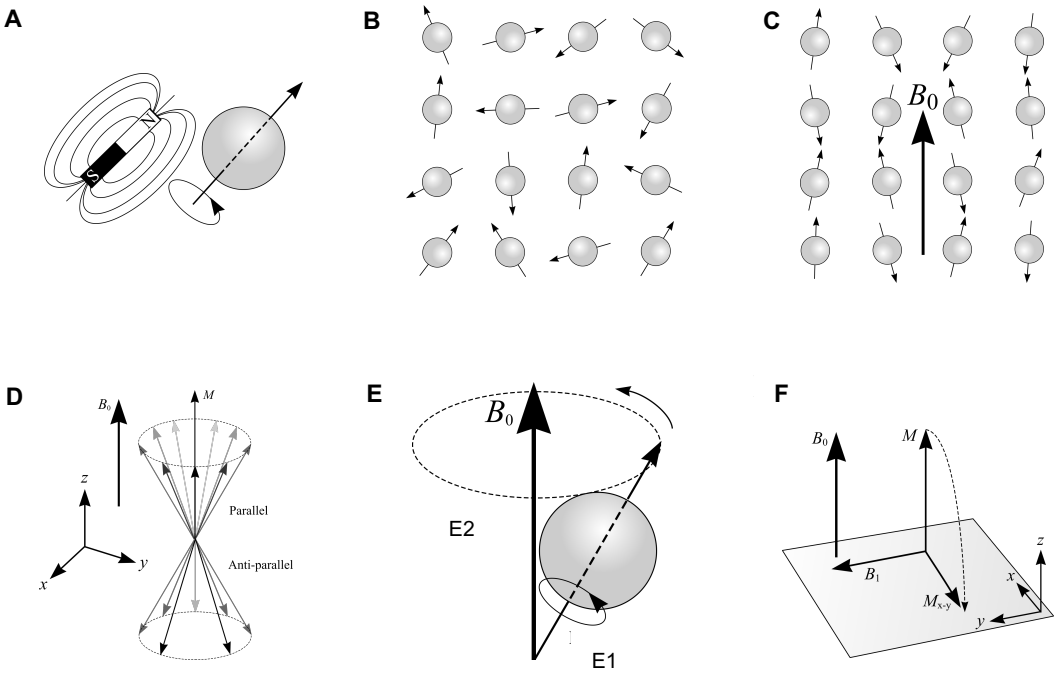


Figure 2.2. MRI physics. **A.** Charged, spinning hydrogen proton creates a magnetic moment. **B.** In the absence of an externally applied magnetic field, protons have random orientations. **C.** When an external magnetic field B_0 is applied the protons align themselves parallel or anti-parallel with respect to B_0. **D.** The net alignment M is oriented along B_0 and the z-axis. **E.** Proton spinning: **E1.** The atom spins in its own axis, **E2.** Precession. **F.** RF pulse produces a second magnetic field M_{x-y} and M is tilted from its original longitudinal z-axis orientation, along the direction of the external magnetic field B_0, into the transverse x-y plane. [Adapted from Puddephat (2010)]

Once the RF pulse is removed, the protons release the absorbed energy and gradually return to their original lower energy state; this is known as relaxation. Relaxation can be measured in two directions: longitudinal relaxation (T1) and transverse relaxation (T2). Longitudinal relaxation (T1) refers to the process in which protons flip back to their original low-energy state parallel to the primary magnetic field B_0 (z-axis), which results in an increase in the longitudinal magnetisation. Plotting the recovery of longitudinal magnetisation over time produces an exponential curve, called the T1 curve. It is difficult to exactly pinpoint the end of longitudinal relaxation. Therefore,

T1 refers to the time taken for longitudinal magnetisation to regrow approximately 63% of its final value. Critically, not all protons return to their original energy state at the same time; different tissues have different rates of T1 relaxation. This allows to create images at a time when the distance between the different tissues T1 relaxation curves is maximal. The result is known as a T1-weighted image, where tissues with a long T1, such as the cerebral spinal fluid (CSF), have low signal intensity and therefore appear dark on the image, whilst those with a short T1 such as white matter (WM) have a high signal intensity and therefore appear bright on the image (Figure 2.3A). Transverse relaxation (T2), on the other hand, occurs when the protons that were in-phase begin to de-phase in the transversal plane (x-y plane), which results in a reduction in transverse magnetisation. Similarly to T1 relaxation, it is possible to plot T2 relaxation over time. This time, T2 is defined as the time that it takes the transverse magnetization to decrease to 37% of its starting value. Contrary to T1, in T2-weighted images, tissues with a long T2 such as CSF, appear brighter, and tissues with a shorter T2 such as WM appear darker (Figure 2.3B).

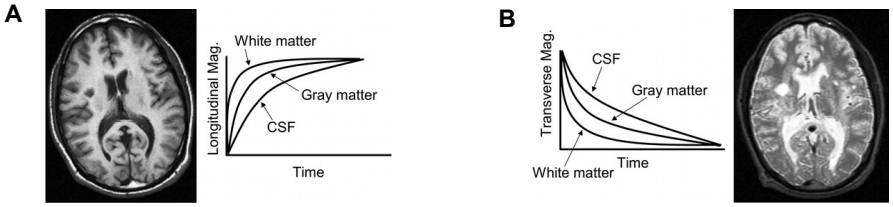


Figure 2.3. T1 and T2 relaxation curves and respective images. **A.** T1-weighted image. **B.** T2-weighted image. [Adapted from (Puddephat (2010)]

2.2.1. Image formation

In order to build a 3-dimensional image (3D), it is necessary to identify the location within the brain from which the RF signal was emitted. This is done by superimposing magnetic field gradients on the otherwise homogeneous external magnetic field B_0. This is achieved using three separate magnetic field gradients, one for each phase of image formation: 1) slice-selection, 2) phase-encoding and 3) frequency-encoding.

2.2.1.1. Slice-selection

Slice localisation is achieved by using gradient coils that generate a gradient field. This means

that different cross sections of the brain will experience a magnetic field of different strength. Accordingly, protons will precess at different frequencies depending on their position along the gradient. Therefore, when the RF pulse is applied, only the protons precessing at the same frequency as the RF pulse will 'flip' into the transverse plane. As a result, only the signal from the protons in this location will be picked up by the receiver coil. This allows a given slice to be selected along the z-axis, with its thickness determined by the strength of the superimposed magnetic field gradient. Once a given slice is selectively excited, the signals arising from each slice element – pixel – within that section need to be spatially encoded. This is achieved using phase- and frequency-encoding gradients.

2.2.1.2. Frequency encoding

During frequency encoding, a different magnetic field gradient is superimposed upon B_0 such that the precessional frequency of protons in the already selected slice is graded along the x-axis. Application of an RF pulse and subsequent recording of the MR signal results in spatial encoding along the x-axis, reflecting the interference pattern formed by the different frequencies along the x-axis.

2.2.1.3. Phase encoding

Phase encoding requires the application of a magnetic field gradient superimposed upon B_0 and is used to account for the alterations in phases that differ along the gradient applied in frequency encoding. Here, the gradient is applied along the y-plane orthogonal to those used in slice selection and frequency encoding. A pulse sequence is then repeatedly applied with only the phase encoding gradient changing, with field strength declining to zero and then increasing back to its original amplitude. The number of times the pulse sequence is repeated is equal to the number of pixels in the subsequent image matrix generated.

Repeating phase and frequency encoding for each slice along the z-axis, corresponding information is collected for each volume-element, or voxel, the size of which is governed by the slice selection gradient. A Fourier transformation can then be used to generate a signal intensity for each, based on the phase and frequency information gathered during encoding, which in turn

can be converted into intensities on a grey scale forming a 3D volumetric image, the resolution of which depends on the voxel size.

2.2.2. MRI acquisition parameters

Site 1: Chengdu University, China

High-resolution 3D T1-weighted images were acquired on a 3 T General Electric MRI scanner (Milwaukee, WI, USA) at the Huaxi MR Research Centre in Chengdu. Images were acquired using a spoiled gradient-recalled acquisition (SPGR) sequence with the following parameters: time TR=8.5ms, TE=3.4ms, flip angle=12°, voxel size=0.47x0.47x1mm, matrix=512x512x156.

Site 2: King's College London, England

High-resolution 3D T1-weighted images were acquired on a 3 T General Electric MRI scanner (Milwaukee, WI, USA) at the Maudsley Hospital, London. Images were acquired using a SPGR sequence with the following parameters: TR=6.9ms, TE=2.8ms, flip angle=18°, voxel size=1.02x1.02x1.2mm, matrix=256x256x166.

Site 3: Santander University A, Spain

High-resolution 3D T1-weighted images were acquired on a 3T Philips Medical Systems MRI scanner (Achieva, Best, The Netherlands) at the Hospital Marques of Valdecilla, Santander, Spain. Images were acquired using a spoiled gradient-recalled acquisition (SPGR) sequence with the following parameters: TR=8.2ms, TE=3.7ms, flip angle=8°, voxel size=0.94x0.94x1mm, matrix=256x256x160.

Site 4: Santander University B, Spain

High-resolution 3D T1-weighted images were acquired on a 1.5T General Electric MRI scanner (Milwaukee, WI, USA) at the University Hospital Marques of Valdecilla, Santander, Spain. Images were acquired using a SPGR sequence with the following parameters: TR=24ms, TE=5ms, flip angle=45°, voxel size=1.02x1.02x1.5mm, matrix=256x256x124.

Site 5: Utrecht University, The Netherlands

High-resolution 3D T1-weighted images were acquired on a Philips 1.5T Achieva MRI scanner (Philips Medical Systems, Best, The Netherlands) at the University Medical Center Utrecht. Images were acquired using a SPGR sequence with the following parameters: TR=30 ms, TE=4.6ms; flip angle=30°, voxel size=1x1x1.2mm, matrix=256x150x150.

2.2.3. Preprocessing

Three approaches were used to preprocess the structural MRI (sMRI) images: voxel-based morphometry (VBM), voxel-based cortical thickness (VBCT) and surface-based morphometry (SBM). All three approaches allow the extraction of neuroanatomical information from T1-weighted images. However, while the first two share some of the main assumptions and data preprocessing, the latter rests on an entirely different approach.

VBM uses deformation fields to identify focal differences in cerebral tissue, either white or grey matter, by comparing different brains on a voxel-by-voxel basis while discounting large scale differences in gross anatomy and position (Ashburner & Friston, 2000). The original purpose of this preprocessing was to generate brain images that would allow comparing at least two groups of brains on a voxel-by-voxel fashion. However, more recent studies are also using the same preprocessed images for machine learning analysis (Nieuwenhuis et al., 2012; Pettersson-Yeo et al., 2013). This section describes the main steps involved in VBM preprocessing as implemented by the Statistical Parametric Mapping (SPM) software (http://www.fil.ion.ucl.ac.uk/spm). More recently, this approach has been extended to VBCT, a voxel-based method to measure cortical thickness, as described in Hutton (2008). The main motivation for this method is that while VBM provides a mixed measure of cortical grey matter including cortical surface area or cortical folding as well as cortical thickness, VBCT selectively investigates cortical atrophy. Consequently, both can be used in combination to build a more complete description of the extent of neuroanatomical alterations (Hutton et al., 2009). Surface-based methods on the other hand, rely on geometrical models that reconstruct the cortical surface from T1-weighted MRI images to quantify different aspects of brain anatomy, including the volume of cortical and subcortical structures as well as cortical thickness. As with VBM and VBCT, there are several different implementations of SBM

(e.g. Brain Visa, CARET, Brain Voyager). In this thesis, SBM was implemented with FreeSurfer (surfer.nmr.mgh.harvard.edu).

Studies using a combination of the different three methods – VBM, VBCT and SBM – have reported different results, which has been attributed to differences in the methods themselves and/or the underling biology being measured with each method (Blankstein, Chen, Mincic, McGrath, & Davis, 2009; Voets et al., 2008). Importantly for machine learning analysis, these different approaches can also result in data with different dimensionalities. While whole-brain voxel-based data contains thousands of voxels, surface-based volumes and thickness are typically used in a region of interest (ROI) approach and therefore have much lower dimensionality. Based on the described above, three sets of data were extracted from each structural image: two voxel-based anatomical measures – 1) voxel-wise grey matter volume (VWGMV) extracted using VBM, 2) voxel-wise cortical thickness (VWCT) extracted using VBCT as well as 3) surface-based regional subcortical and cortical GM volume and cortical thickness extracted using SBM (SB-ROIs). In what follows, a description of the preprocessing for each type of neuroanatomical measure is provided.

2.2.3.1. Voxel-based anatomical measures

Both VWGMV and VWCT are voxel-based measures and therefore share some of the preprocessing steps. Common to both measures, structural images were first reoriented along the anterior-posterior commissure line and set the anterior commissure as the origin of the spatial coordinates to assist with the normalization algorithm. The unified segmentation procedure (Ashburner & Friston, 2005) was then used in combination with the Diffeomorphic Anatomical Registration through the Exponentiated Lie algebra (DARTEL) algorithm (Ashburner, 2007) to segment and subsequently normalise the reoriented structural images. This process consists in first segmenting each image into GM, WM and CSF. This is achieved by assigning each voxel a probability of it being GM, WM and CSF, such that the sum of the tissue probabilities at each voxel adds up to one. The output is a set of three new images for each subject, each one a probability map for a type of tissue in the space of the input data (i.e. native space). The GM and WM segmented partitions were then used to create study specific template derived from the mean of

all subjects, thus allowing a voxel-for-voxel correspondence across the subjects' images. The aim of this procedure is to reshape each brain to a more appropriate and study-specific template, as opposed to a more generic one (e.g. MNI-305) (Ashburner, 2007). This is achieved through a nonlinear registration which allows local areas to stretch and compress with respect to each other. The transformations necessary to match each voxel in the input image to match the template are mapped in a deformation field. Once the images are segmented and the DARTEL template created, the rest of the preprocessing is done using two separate approaches - VBM and VBCT - to extract VWGM and VWCT, respectively.

2.2.3.1.1. Voxel-based morphometry

To create the VWGM maps, the segmented GM partitions were warped to the new study-specific reference space, using each subject-specific deformation field. This creates an image that is in voxel-for-voxel registration with the template. The warped GM partitions were then affine-transformed into MNI space. A further processing step referred to as "modulation" was also used to compensate for the expansions and/or contractions each voxel was subjected to when the deformation field was applied, thus ensuring that the total amount of signal in each voxel was conserved (C. D. Good et al., 2001). Finally, the GM probability maps are smoothed in a process that involves convolving an isotropic kernel across each image such that the intensity in each voxel is a locally weighted average of the signal intensity from a region of surrounding voxels as defined by the size of the kernel (Ashburner & Friston, 2000). There are several reasons for using smoothing. First, it compensates for residual anatomical variability after spatial normalization and renders the data more normally distributed increasing the validity of any statistical parametric test. Second, it is required to comply with the assumptions underlying Gaussian Field theory. Additionally, smoothing also reduces the effective number of statistical comparisons, thus making the correction for multiple comparisons less severe. The value at a voxel in the final modulated smoothed image is interpreted as the volume of GM at that location.

2.2.3.1.2. Voxel-based cortical thickness

The first step of VBCT is to create a cortical thickness map for each subject. This consisted in using the GM, WM and CSF tissue partitions and a nonlinearity matched labelled brain atlas to

first automatically extract the inner and outer cortical GM boundaries followed by the measurement of the distance between the two. This process can be challenging since the cortical sheet is highly folded and with variable thickness. In the method proposed by Hutton et al. (2008) this is addressed by dividing the cortex into sub-layers and consider the thickness of each sublayer separately. This approach has two motivations. First, it has an analogous association to the known layer-wise structure of the cortex. Second, the calculations necessary to estimate the thickness of adjacent layers is similar to that of other mathematical problems and can be solved using Laplace's equation (Jones, Buchbinder, & Aharon, 2000). The resulting VBCT maps contain the cortical thickness values within voxels identified as GM and are saved in the native space of the input images. Once created, each VBCT map was warped to the new study-specific reference space by applying the corresponding subject specific deformation field. The warped images were then modulated and smoothed with Gaussian kernel. The same warps, modulation and smoothing were also applied to a binary mask created from each original VBCT map. The smoothed VBCT maps were divided by the corresponding smoothed mask. The effect of this procedure was to project the Gaussian smoothing kernel applied to the warped images, into the native space of the subject while preserving the cortical thickness value over a region the size of the smoothing kernel. As a final step, the warped and smoothed VBCT maps were affine transformed into MNI space.

2.2.3.2. Surface-based morphometry

FreeSurfer is a widely used method for processing anatomical MRI images (Fischl, 2012). Its 'recon-all' processing stream is a fully automated procedure that takes a T1-weighted image as the input and outputs the segmented image as well as cortical measures such as volumes and surface. The technical details of each stage have been extensively described elsewhere (Dale, Fischl, & Sereno, 1999b; Fischl & Dale, 2000; Fischl, Sereno, & Dale, 1999; X. Han et al., 2006). Briefly, each raw T1-weighted image is first corrected for intensity bias to address the intensity variations due to magnetic field inhomogeneities. Any extra-cerebral voxels are then removed from the resulting normalized intensity images, using a 'skull-striping' procedure. This is followed by the automatic segmentation of the subcortical WM and deep GM volumetric structures (e.g. hippocampus, amygdala, caudate, putamen, ventricles) (Dale et al., 1999b). The segmented WM

volume is then used to derive a tessellated surface representing the grey/white matter boundary (inner surface, also referred to as WM surface), which is automatically corrected for topology defects and expanded to model the pial–grey boundary (outer surface, also referred to as pial surface). Each surface can be described as a mesh of vertexes, connected by edges, that form a mesh of tessellated triangles along the cortical mantle. Once the cortical models are completed, a number of deformable procedures are performed for further data processing including surface inflation to allow the measurement of hidden sulci and registration to a spherical atlas which is based on the individual cortical folding patterns to match cortical geometry across subjects. Finally, each spherical model is parcellated based on a spherical in-built atlas [Desikan-Killiany atlas (Desikan et al., 2006)] to extract morphometric measurements (e.g. cortical thickness and surface area) for specific cortical regions (e.g. fusiform gyrus, middle temporal cortex, superior frontal gyrus) (Fischl et al., 1999). Based on this procedure, several cortical metrics can be estimated by the use of different overlays that assign a value for each metric (e.g. thickness, volume, curvature) for each vertex. For example, for each vertex, cortical thickness is calculated as the average of the distance from the WM surface to the closest point on the pial surface and from that vertex back to the closest point on the WM surface (Fischl & Dale, 2000).

Within this thesis, three SB-ROIs measures were used: 1) volumes of subcortical regions, 2) thickness and 3) volume of cortical regions from each hemisphere. Most ROIs were measures in both hemispheres separately, totalling 169 ROIs. A complete list of all the regions included in the data analysis is shown in Table 2.1.

Table 2.1. List of cortical and subcortical brain regions extracted with FreeSurfer.

Cortical structures	Subcortical structures
Banks of superior temporal sulcus	Third ventricle
Caudal anterior cingulate	Fourth ventricle
Caudal middle frontal gyrus	Brainstem
Cuneus cortex	Corpus callosum anterior
Entorhinal cortex	Corpus callosum central
Fusiform gyrus	Corpus callosum midanterior
Inferior parietal cortex	Corpus callosum midposterior
Inferior temporal gyrus	Corpus callosum posterior
Isthmus of cingulate cortex	CSF
Lateral occipital cortex	Accumbens
Lateral orbitofrontal cortex	Amygdala
Lingual gyrus	Caudate
Medial orbitofrontal cortex	Cerebellum cortex
Middle temporal gyrus	Cerebellum white matter
Parahippocampal gyrus	Hippocampus
Paracentral sulcus	Inferior lateral ventricle
Frontal operculum	Putamen
Orbital operculum	Lateral ventricle
Triangular part of inferior frontal gyrus	Pallidum
Pericalcarine cortex	Thalamus proper
Postcentral gyrus	Ventral DC
Posterior cingulate cortex	
Precentral gyrus	
Precuneus cortex	
Rostral anterior cingulate cortex	
Rostral middle frontal gyrus	
Superior frontal gyrus	
Superior parietal gyrus	
Superior temporal gyrus	
Supramarginal gyrus	
Frontal pole	
Temporal pole	
Transverse temporal cortex	
Insula	

2.3. Data Analysis

2.3.1. Univariate analysis

2.3.1.1. Voxel-based morphometry

Group-level differences between FEP and HC on VWGMV and VWCT were estimated using the general linear model (Frackowiak, 2004). According to this framework, the data can be described in terms of effects of interest, confounds of no interest, and residual error. Statistical analysis is performed by fitting a pre-defined model to the data to estimate the contribution, i.e. parameter weights, of each variable of interest to the observed data. Standard parametric tests (t-test and

F-test) can then be applied to the estimated parameters to identify differences between effects of interest at each voxel. The results are then used to generate a statistical parametric map (SPM), where each voxel is assigned a t-statistic. Given the mass-univariate nature of the approach, a correction for multiple comparisons based on Gaussian random field (GRF) theory is also to minimise the risk of false positives (Worsley et al., 1996).

2.3.1.2. Surface-based morphometry

Differences in mean for each SB-ROI between FEP and HC were analysed with an independent-sample t-test as implemented in SPSS 24.0 using a statistical threshold of $p<0.05$ and additional Bonferroni correction for multiple comparisons.

2.3.2. Machine learning

The supervised machine learning pipeline for a binary task can be divided into two main stages: training and testing. During the training stage, a machine learning algorithm identifies the features that best distinguish the two groups. In the testing phase, a previously unseen subject is assigned to one of the groups based on the decision function the algorithm learned during training. Within this general framework, there are a multitude of machine learning algorithms that can be applied. This thesis focuses on the application of a deep learning algorithm known as deep neural network. However, it is considered to good practice to test different approaches for the same task. Therefore, three additional and well-established approaches were also used: 1) K-nearest neighbours (KNN), 3) logistic regression (LR) and 3) support vector machine (SVM). These machine learning algorithms were chosen based on their increasing order of complexity: KNN is a straightforward algorithm often used as a benchmark, whilst deep learning can be more powerful at the expense of transparency; and popularity: SVM and LR and among the most commonly used techniques used in previous studies. The implementation of each one if these algorithms followed the same pipeline: 1) dealing with confounding variables, 2) feature extraction and dimensionality reduction 3) scaling, 4) model training and finally 5) model evaluation (Figure 2.4). The following subsections provide the details of the strategies and techniques used to build this machine learning pipeline.

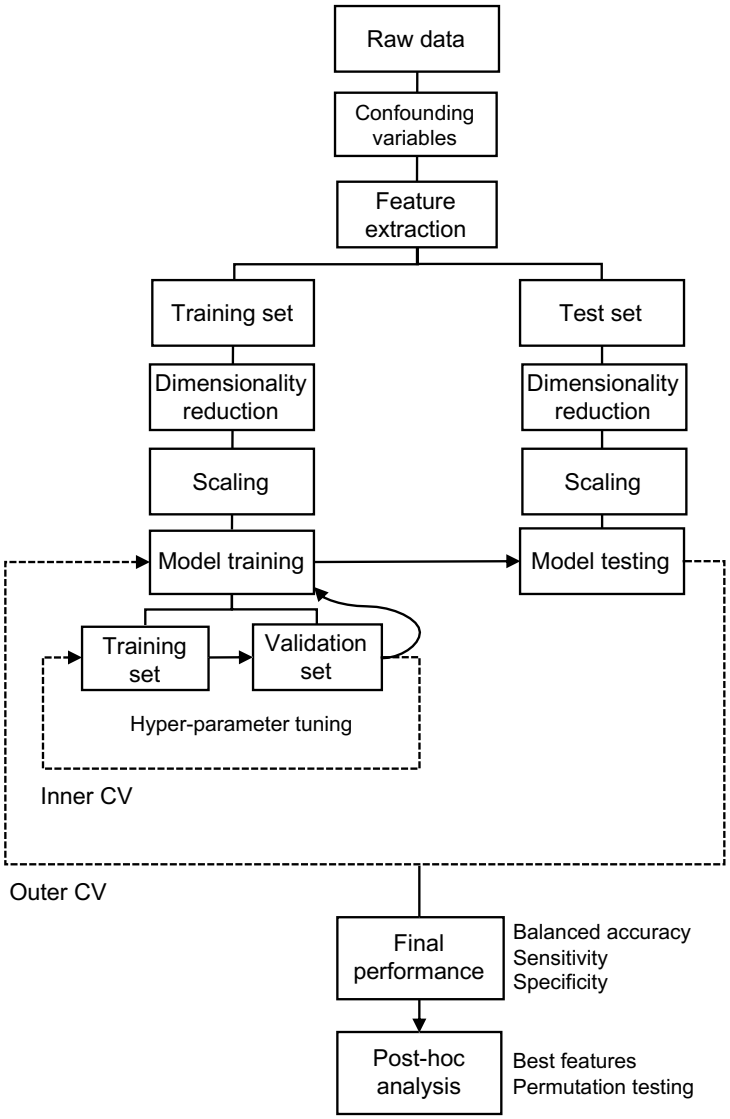


Figure 2.4. Summary of the machine learning pipeline implemented in this thesis.

2.3.2.1. Confounding variables

Given the well-established effect of sex and age on brain neuroanatomy (Luders et al., 2004; Pina-Camacho et al., 2016), these two demographic variables were treated as possible confounders. In order to mitigate their effect, FEP and HC groups from each site were matched for sex and age. To maximize the use of the data made available, matching was carried out by taking the group with smallest sample size and randomly selecting participants from the other group according to age (+/- 5 years) and sex. For all sites, the FEP:HC matching ratio was 1:1, except for site 4 where the ratio was 2:1. Due to this imbalance in site 4, balanced accuracy was used as the performance metric of choice and all machine learning algorithms were trained using a stratified CV to ensure a similar FEP/HC ratio across all iterations of the CV. The final sample

sizes for each site are shown in Table 2.1. and Chapter 5.

2.3.2.2. Feature extraction and dimensionality reduction

The VWGMV, VWCT and SB-ROIs data extracted using the procedures explained in section 2.2.3. were used as input features for the diagnostic classification of FEP and HC. The preprocessed VWGMV and VWCT maps contain several thousands of voxels, i.e. they are very high-dimensional. The deep learning architecture used in this work is not designed to handle whole-brain voxel-level data as this would result in an unfeasible number of parameters to estimate during training (see section 2.3.2.4.2. Deep neural networks). A possible solution to this issue is to extract a new set of features that compress the information in the original data into a substantially smaller number of features. This procedure is known as dimensionality reduction. Therefore, to mitigate the likelihood of overfitting and alleviate computational requirements, the dimensionality of VWGMV and VWCT maps was reduced via principal components analysis. Two of the alternative methods used in this work – LR and SVM – are regularized methods and therefore do not need this step. However, PCA was also used in combination with these methods to facilitate comparison between all approaches as well as to alleviate computational requirements during training.

Principal component analysis (PCA) is a well-established unsupervised method for dimensionality reduction that has been widely used in neuroimaging (Mwangi, Tian, & Soares, 2014) and which technical details have been extensively reported (Lever, Krzywinski, & Altman, 2017). Using this approach, the original and likely correlated features are transformed into a set of values of linearly uncorrelated variables called principal components. Briefly, this is achieved by projecting the original data into a new coordinate system where the first axis, called the first principal axis, corresponds to the direction along which the data varies the most; the second axis, called the second principal axis, corresponds to the direction along which the data varies the most after the first direction; and so on. The first principal component is the projection of the original data onto first principal axis and captures the greatest amount of the variance in the data. The second principal component is the projection of the original data onto second principal axis and explains the greatest amount of the variance in the data that is not captured by the first principal

component. Data is thus decomposed such that each subsequent principal component explains the greatest amount of variance possible under the constraint that it is orthogonal to the preceding principal components. Dimensionality reduction can therefore be attained by selecting only the principal components that capture the most variability of the original data. In this thesis, PCA was implemented with the function PCA from the module decomposition from the Scikit-Learn library (sklearn, version 0.20) for python 3.5.

2.3.2.3. Scaling

To model the data correctly and effectively, most machine learning algorithms require the data to be on the same scale. This is because if a feature's variance is orders of magnitude greater than the variance of other features, that particular feature might dominate the others in the dataset. There are several possible solutions to avoid this issue, collectively known as feature scaling. In this thesis, data was transformed such that the distribution of each feature resembles a standard normal distribution with mean=0 and variance=1; this is known as standardization or z-score normalization. Each normalized value z_{x_i} is calculated by taking each data point x_i, subtracting the mean $\overline{X}$ and then dividing by the standard deviation (SD) of the same feature:

$$z_{x_i} = \frac{(\overline{X}_{Feature_A} - x_i)}{SD_{Feature\ A}} \tag{2.1}$$

In this work, scaling was implemented using the function StandardScaler from sklearn (version 0.20, Pedregosa et al. (2011)) for python 3.5.

2.3.2.4. Model training

2.3.2.4.1. Cross-validation

The training and testing of each machine learning algorithm was implemented via nested stratified 10-fold cross-validation (CV). This involves dividing the total data into 10 groups, training the model in 9 groups and use the left-out group for testing. This is done iteratively, using a different group for testing each turn, until all groups have been used for testing. Given the 2:1 ratio of FEP and HC in site 4, the ratio FEP:HC was kept consistent across iterations of the CV to avoid a disproportionate number of patients and controls, which could affect training. Critically, each one

of the machine learning algorithms used in this work relies on the specification of different hyperparameters. This choice of values for these hyperparameters was done via nested CV. This creates a second, inner, CV inside the already defined primary, outer, CV. At each iteration of the outer CV, the training set is further divided into training and validation sets, where different possible values for the hyperparameters are fitted to the training set and tested in the validation set. The hyperparameters with the best performance in the validation set are then used to fit the model to the training set as defined by the outer CV. The final performance is subsequently estimated by averaging the performance in the test set across all outer CV iterations.

In order to ensure the independence between training and test sets necessary for an accurate measure of generalizability of the trained model, PCA and normalization were implemented as part of the CV scheme. For PCA, this was done first extracting the minimum number of principal components whilst retaining cumulative 90% of the variance from the data in the training set only. The VWGMV/VWCT maps were then projected onto the resulting principal components and the resulting values were used for training the classifier. The VWGMV/VWCT maps from the test set were projected onto the same components derived from the training set and the resulting values were used for testing the classifier. Likewise, the scaling procedure was implemented by first estimating the mean and standard-deviation separately for each feature in the training set only. These were subsequentially used to scale both the training and test sets. Both PCA and scaling procedures were done iteratively for each fold of the CV.

2.3.2.4.2. Deep neural networks

The flexibility inherent to deep learning models has propelled a vast family of possible architectures, each one built for a specific purpose. Amongst all the possible architectures, deep neural networks (DNN) (or multilayer perceptrons) emerges as the simplest and most straightforward application of deep learning. This section provides an overview of the main elements of DNN: structure, training, regularization ang hyperparameter tuning.

2.3.2.4.2.1. Structure

DNNs are organized in a layer-wise structure. Each layer comprises a set of neurons (also known

as units or artificial neurons), loosely inspired by the biological neuron (Figure 2.5). Each neuron is connected to the neurons in adjacent layers to create a network, akin to a network of biological neurons. Connections are represented by weights w that reflect the strength and direction (excitatory or inhibitory) between two neurons.

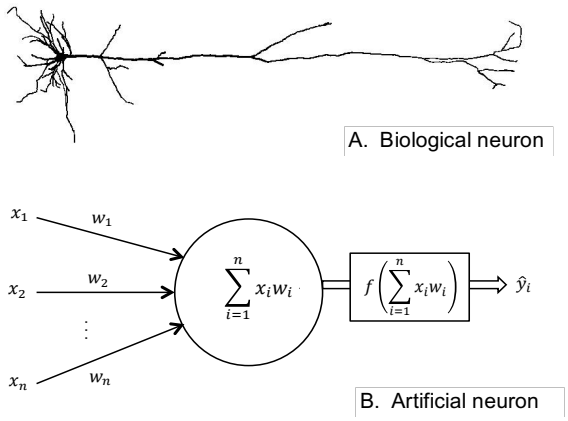
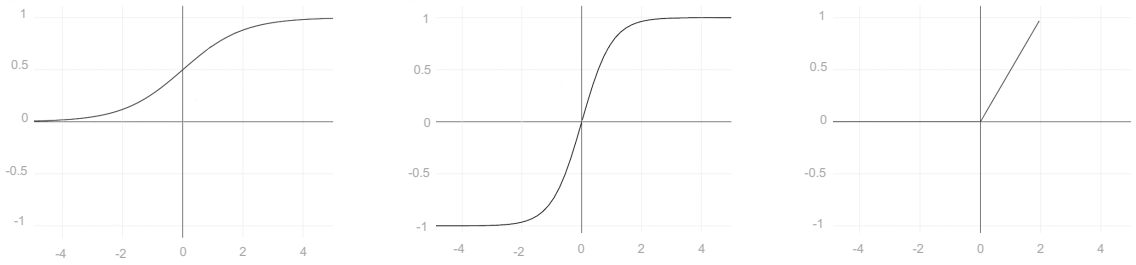


Figure 2.5. A. Biological neuron **B.** Artificial neuron.

Each neuron transforms the incoming data x by calculating the weighted sum of the output of the neurons in the previous layer; then passing it through a nonlinear function f to derive the output for that neuron. There are several nonlinear functions, also known as activation functions, that can be used; some of the most common ones include the rectified linear unit (ReLU), hyperbolic tangent activation (tanh) or the sigmoid function (Figure 2.6).



$$f(x) = \frac{1}{1 + e^{-x}} \qquad f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad f(x) = max(0, x)$$

Figure 2.6. Example of commonly used activation functions: sigmoid (left), tanh (centre) and ReLU (right).

The first layer of the network corresponds to the input layer where data is entered into the model (Figure 2.6). The last layer is the output layer. In a classification task, the number of neurons in the output layer corresponds to the number of classes. Here, a special nonlinear function is typically used to yield the probability of a given subject belonging to each class; this nonlinear function is called softmax function. The layers between the input layer and the output layer are referred to as hidden layers, and their number represents the depth of the model (hence the term 'deep' learning). In a typical network, all neurons in one layer are connected to all neurons in adjacent layers; this is known as a fully-connected network. The consecutive nonlinear transformations and propagation of information form the input through the hidden layers until it reaches the output layer, is known as forward propagation.

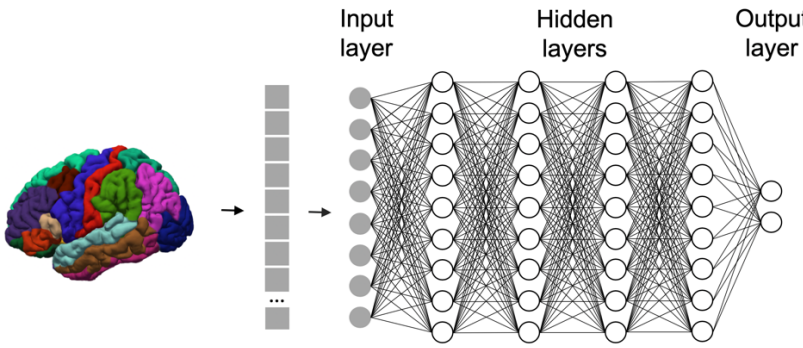


Figure 2.7. Exemplar application of a DNN to neuroimaging data. SM-ROIs are transformed into a 1D vector and used as input data for the first layer of a fully-connected DNN.

In order to fully-describe the network before training, the number of layers and neurons as well other hyperparameters need to be specified. In this thesis, this was done via automatic hyperparameter tuning, discussed in section 2.3.2.4.3.4.

2.3.2.4.2.2. Training

In a typical DNN, training consists of an iterative process of adjustment of the weights between the neurons within the network, much like a human brain learns through the fine-tuning of connections between neurons (Bengio, 2009). These weights can be thought of as 'knobs' that determine the relationship between the input data and the network's output (LeCun et al., 2015). The main aim of training is to tweak these 'knobs' in such a way that maximizes the strength of

the relationship between input and output. Weights can be learned by framing training as an optimization problem: find the network's weights that minimize the difference between prediction and true target (i.e. error). There are several ways in which this optimization can be implemented. The most commonly used are Gradient Descendent (GD) based optimizers (such as vanilla GD or Adam (Kingma & Ba, 2014)). This section provides an intuitive description of GD (a detailed technical explanation can be found in Goodfellow (2016)).

An intuitive way to describing GD is to think of the optimization task as the search for the lowest point in a chain of mountains. In this analogy, the loss function can be thought of as a chain of mountains and valleys, in which every point along the chain is associated with a possible set of values of the network's weights, and where the deeper the valley, the smaller the error; this is usually referred to as the loss landscape (Figure 2.7). In turn, optimization through GD can be thought of as a way to navigate through this landscape to the bottom of the deepest valley, also known as global minima. At the beginning of training, the weights of the DNN are initialized at random. A training observation (e.g. data from one participant) is then entered into the input layer and the information forward propagated through the network until it reaches the output layer, where the network outputs a prediction value and the error is calculated using a loss function. In the analogy described above, the value of this error can be thought of as a random location in the loss landscape, from which the optimizer will navigate in search for the global minima (point A in Figure 2.7).

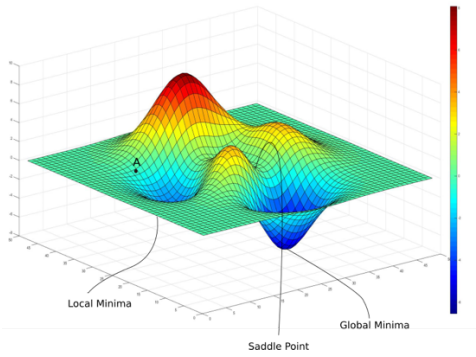


Figure 2.8. Loss landscape. The random initiation of the DNN results in an error of magnitude equivalent to a random point in the landscape (point A). The lower the 'valley', the smaller the error of the loss function. Ideally, at the end of training, the loss function will have reached the global minima, i.e., the model has a

very small error.

From here, the GD algorithm checks which direction, out of all possible options, results in the steepest decline, i.e., a larger decrease in the value of the loss function; this is formally known as the gradient: its direction reflects the direction with the steepest descent, and its magnitude indicates how steep the descent is. Once the direction of the steepest descend has been established, the size of the step towards that direction needs to be determined; this is known as learning rate. Once the gradient and learning rate are determined, the optimizer takes its first step. Formally, this is translated in a set of updated weights. Once a further data observation is passed thought the network, and a new value for the loss function – in principle of smaller amount – is generated. This corresponds to a new position in the loss landscape. At this new position, the gradient is recomputed, and the optimizer takes another step towards the minima. This is done iteratively and, as the optimizer approaches the minima, the contour of the function, i.e. the ridges of the valley, become almost flat. This will result in a very small gradient, i.e. there is no more 'downwards' to go from here, and a minima has been reached. Critically, the size of the step – learning rate – is typically adjusted as the optimizer gets closer to the minima. This is because, while it may be useful to take larger steps at the beginning, this might result in overshooting the minima once the optimizer gets closer to it. Therefore, the size of the learning rate is typically reduced along the iterations. This is specified by another hyperparameter of GD, known as learning rate decay. Importantly, it is not trivial to calculate the gradient of the weights of the network's hidden layers at each iteration of the GD. It was only in the 1980s that an efficient method of computing gradients was developed, known as "backward propagation of errors" or *backpropagation* (LeCun et al., 2015; Schmidhuber, 2015; Werbos, 1982).

The type of GD can vary depending on the number of training samples used to calculate the error. There are three main variations of the GD: stochastic, batch and mini-batch. In stochastic gradient descent, the error and respective updates are computed for each observation in the training dataset. Perhaps the most intuitive of the GD variations, it can result in a noisy gradient signal, making it hard for the algorithm to settle on a minimum value. Batch gradient descent on the other hand, considers the error of all training samples before updating the weights. Here, the more

stable gradient signal may converge prematurely, thus resulting in a less optimal network's weights. Finally, mini-batch gradient descent splits the training data into small batches – mini-batches – that are used to calculate the error and update the model coefficients. This approach represents a reasonable trade-off between the two other methods, and as such is the most common form of GD and the one used in this thesis. Similarly to the other hyperparameters, batch size also needs to be determined a priori or tuned as part of the machine learning pipeline.

2.3.2.4.2.3. Regularization

Due to the large number of connections between neurons, the training of DNNs involves the estimation of a considerable number of parameters, i.e. weights. This can lead to the model learning particular fluctuations in the training data that only work in the training set, i.e. overfitting. Therefore, modern DNNs try to minimize this risk by applying different strategies, collectively known as regularization. In this thesis two forms of regularization were used: L2 norm and drop-out. L2 norm involves penalizing models with very high weights. By forcing weights to remain low, the network becomes less dependent on the training data (i.e. performance does not rely heavily on a particular set of weights) and can better generalize to unseen data (Nowlan & Hinton, 1992). Dropout, on the other hand, consists of temporarily removing a random number of neurons and their respective incoming and outgoing connections from the network during training. This results in the extraction of different sets of features that can independently produce a useful output, which in turn has the effect of enhancing generalizability (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014).

2.3.2.4.3.4. Model specification and hyperparameter optimization

DNN models rely on the specification of several architectural and learning hyperparameters. A detailed description of how each hyperparameter was specified is provided in Chapter 5 and 6. In brief, each layer was initialized via Glorot initialization (normal distribution) (Glorot & Bengio, 2010). In the output layer, classification was performed by a softmax function. A mini-batch of 8 training samples was used for the VWGMV and VWCT feature sets, while a mini-batch of 128 was used for SM-ROIs. The number of layers and neurons at each layer, optimizer, learning rate, learning rate decay, activation function, epoch, L2 norm and drop-out rate were optimized through

a nested CV from a range of possible values defined *a priori*. A more detailed description of hyperparameter tuning is given in Chapter 5 and 6.

2.3.2.4.3. Traditional machine learning algorithms

2.3.2.4.3.1. K-nearest neighbours

KNN is a straightforward algorithm often used for its ease of implementation and interpretation. Due its simplicity it is often used in the literature as a benchmark for more complex algorithms (Jain, Duin, & Jianchang Mao, 2000). KNN belongs to a special type of learning known as 'lazy learning'. Whilst most algorithms learn an optimal function that map features and target variable during a training phase, which is then tested in the test set, KNN does not rely on an explicit training phase; instead, it simply stores the entire training data in memory, which is subsequently used as "knowledge" to make predictions of unseen data. Once the training data has been stored, the distance between the new observation in the test set and each observation in the training set is estimated. This distance works as a proxy for similarity and is typically calculated using the Euclidean distance, which estimates the distance between two points by calculating the length of the straight-line between them, such that the larger the distance, the farther and less similar they are to each other. Formally, the Euclidean distance between two data points $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ is given by:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (2.2)$$

Once all distances have been estimated, the algorithm identifies a set of training observations that are closest, i.e. neighbours, to the new unseen observation. The size of this set, i.e. neighbourhood, is given by the hyperparameter k. In case of classification, the unseen observation will be assigned to the majority class among the k observations. The appropriate choice of k has significant impact on the performance of KNN algorithm. A small k forces the algorithm to ignore the overall distribution of the training data. As a result, it provides a flexible fit, with low bias but high variance and a jagged decision boundary. On the other hand, a higher k

considers more observations and hence is more resilient to outliers. This results in a smoother decision boundary, with lower variance but increased bias.

2.3.2.4.3.2. Logistic regression

Logistic regression is one the simplest yet more powerful machine learning algorithms. Either in its simplest form or in more complex variations, logistic regression has been extensively used in the literature. Contrary to linear regression where the outcome is a real number in a continues scale, logistic regression models the probability of belonging to a certain class. Formally, for a binary classification problem with labels A and B, logistic regression models the log odds of belonging to class A as opposed to class B, given the observed predictor variables:

$$\log\left(\frac{P(y_i=A\,|x_{i1},x_{i2},\ldots,x_{in})}{P(y_i=B\,|x_{i1},x_{i2},\ldots,x_{in})}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \quad (2.3)$$

Since $P(y_i = B\,|x_{i1}, x_{i2}, \ldots, x_{in}) = 1 - P(y_i = A\,|x_{i1}, x_{i2}, \ldots, x_{in})$, the probability of belonging to class A is given by:

$$P(y_i = A\,|x_{i1}, x_{i2}, \ldots, x_{in}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}}} \quad (2.4)$$

As a linear method, logistic regression finds the optimal linear combination of the input features and predicts an output value by assigning each feature a coefficient or weight learned from the training data. Once learned, these weights are used to output the probability of the input data belonging to a default class (e.g. HC), which can then be converted to a binary prediction, i.e. if the probability of belonging to class A is higher than 0.5, the input data is assigned to this class, otherwise it is assigned to class B. There are different approaches to learn the optimal weights during training (e.g. maximum likelihood estimation, stochastic gradient descent). In this thesis, weight optimization was implemented via stochastic gradient descent (GD), as described in section 2.3.2.4.2.2.

Logistic regression can be combined with regularization strategies to mitigate the likelihood of overfitting. Elastic net is a commonly used method that consists in the combination of two regularization techniques: L1 norm from LASSO (Least Absolute Shrinkage and Selection

Operator) and L2 norm from ridge regression. Both penalize non-zero coefficients. However, this is done differently for each technique. Briefly, the L1 penalty discards features that contribute most to the error by shrinking their weights to zero, effectively working as a feature selector. The reduced number of features reduces model complexity and thus help prevent overfitting. Conversely, the L2 penalty retains all variables, whilst forcing their weights to be low. This will result in a model less reliant on a specific group of features from the training set, and therefore more likely to generalize to new data. From here it follows that, while both methods allow to use correlated predictors, they solve the issue of multicollinearity differently. From the same group of highly correlated features, L1 norm selects one while the rest are removed, whereas L2 norm assigns all features similar weights. Therefore, in the context of neuroimaging data, where high-dimensional data tends to be highly correlated and the effect of each feature is thought to be subtle, the former may result in unwanted loss of information, while the latter will only mitigate model complexity as the number of features is kept the same. Elastic net can therefore be used to combine both penalties and overcome each technique's shortcomings.

2.3.2.4.3.3. Support vector machine

Support vector machine (SVM) is one of the most widely used classifiers in the machine learning, including in psychiatric neuroimaging (Orrù et al., 2012). The technical details of SVM have been extensively reported (Vapnik, 1995). Briefly, SVM try to separate observations into classes using a decision boundary in a high-dimensional space. The separation boundary can then be used to classify unknown observations. In a high-dimensional space, the separation between the two classes can be visualized as a hyperplane as defined by:

$$\mathrm{w}.\mathrm{x} + \mathrm{b} = 0 \tag{2.5}$$

where x represents the input features, w the weight vector and b the bias. However, there are many possible hyperplanes that can separate the two classes. The assumption here is that finding a hyperplane that has the maximum distance, or margin d, between the most difficult data points of either class to classify will maximise generalisability. Formally, the margin can be defined as distance between the points closest to the hyperplane:

$$d = \frac{2}{||w||} \tag{2.6}$$

These data points are the nearest observations of either class to the hyperplane and are collectively known as the support vectors (Figure 2.9).

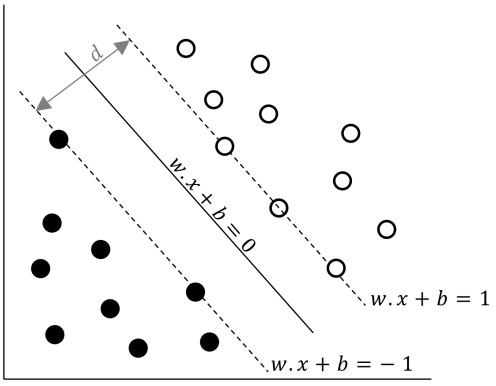


Figure 2.9. Hyperplane and support vectors.

The optimal hyperplane can therefore be obtained through an optimization problem where the aim is to maximize d (which is equivalent to minimizing $||w||$). Once obtained, the classification problem can be solved intuitively by assigning an observation to one of two classes (1 or -1), according to the sign of the equation 2.7 as follows:

$$\mathrm{f(x) = sign(w.x + b)} \tag{2.7}$$

such that, if $\mathrm{f(x)} \geq 0$ the participant is classified as class +1 (e.g. patients) otherwise it is classified as class -1 (e.g. controls).

Since real-world data might not be perfectly separated with a hyperplane, SVM includes a hyperparameter that allows some data points in the training data to violate the separating hyperplane whilst still trying to maximize the margin between data points of different classes. This trade-off is regulated by the soft-margin or C hyperparameter. For C=0 no violation is allowed; this is known as a hard margin classifier. For large values of C, a smaller-margin hyperplane is chosen if it results in all the training points being classified correctly. Conversely, a very small value of C will cause the algorithm to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more data points. In addition to the hyperparameter C, the SVM also

allows for an additional regularization penalty. This is typically either the L1 or L2 discussed in the previous section. While L1 aims to reduce the number of features by assigning a weight of zero to a subset of less important features, L2 keeps all features but forces their weight to remain low. In this thesis, the default L2 regularization strategy was used as implemented in the SVM module in scikit learn library (Pedregosa et al., 2011).

When groups cannot be separated with a linear decision boundary (even with the aid of a soft margin), SVM allows the use of kernels - a similarity function over pairs of data points in their raw representation - that transform the initial features space into a new higher-dimensional space which may allow the distinction of initially nonlinearity separable classes. In this thesis, a linear kernel was used to contrast with the characteristic nonlinear approach of deep learning. This consists of transforming the initial feature space by computing the dot product between each pair of observations.

2.3.2.5. Model evaluation

2.3.2.5.1. Performance metrics

Balanced accuracy was used as the main outcome of interest, along with sensitivity and specificity. Below is a summary of the definition of each metric and how they were calculated.

Balanced accuracy: average between the accuracies for each class or the average between the sensitivity and the specificity.

$$Balanced\ Acc = \frac{\frac{True\ positive}{True\ positive+False\ negative} + \frac{True\ negative}{True\ negative+False\ positive}}{2} \quad (2.8)$$

Sensitivity: proportion of FEP patients correctly identified.

$$\text{Sensitivity} = \frac{True\ positive}{True\ positive+False\ negative} \quad (2.9)$$

Specificity: proportion of HC correctly identified.

$$\text{Specificity} = \frac{True\ negative}{True\ negative+False\ positive} \quad (2.10)$$

2.3.2.5.2. Significance testing

Testing for statistical significance is considered an important step, especially when dealing with small sample sizes, to minimize the risk of over-optimistic conclusions. Significance testing is usually carried out using permutation testing. In essence, this method measures the likelihood that the model's performance would be observed by chance. For a supervised algorithm, this is estimated by first randomly shuffling the target variables for all subjects. This is was done 1000 times such that any statistical relationship between the target variable and the input features was lost. At each permutation, the same previously trained model was applied to the input features and the corresponding randomly assigned labels. Critically, the same stratified cross-validation scheme used during the training of the model was also used here, to ensure a balanced ratio between the two labels in each fold. The balanced accuracy was then calculated for each permutation, resulting in a statistical distribution of balanced accuracy which reflects the null hypothesis that the model behaves by chance. The number of times the performance was greater than or equal to the original performance was then divided by the number of permutations (i.e. 1000) to estimate a p-value (P. Good, 1994). In this thesis, each models' final balanced accuracy was tested for significance and a p-value lower than 0.05 was considered statically significant.

Chapter 3

Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis

This chapter is based on the paper entitled Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis currently under review.

Vieira, S., Gong, Q., Scarpazza, C., Lui, S., Huang, X., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-Garcia, V., Setién-Suero, E., Scheepers, F., van Hareni, N. E. M., Kahn, R., Reis Marques, T., Ciufolini, S., Di Forti, M., Murray, R. M., David, A., Dazzan, P., McGuire, P. & Mechelli, A. (Accepted/In press). Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis. *Psychological medicine*.

3.1. Introduction

Neuroanatomical abnormalities in schizophrenia have been well documented for the past three decades (Glahn et al. 2008; Bora et al. 2011). While the initial research was performed in patients with long-term schizophrenia (Ellison-Wright et al. 2008), more recent studies have focussed on individuals in the early stages of the illness, when the effects of chronicity (Olabi et al. 2011; Vita et al. 2012) and antipsychotic medication (Radua et al. 2012; Vita et al. 2015; Shah et al. 2017) are minimal. The results of these studies, however, tend to be inconsistent from one investigation to another (Radua et al. 2012; Gao et al. 2017; Shah et al. 2017). For example, reports of insular abnormalities have been heterogeneous, with some studies reporting increased (Salgado-Pineda et al. 2003; Ren et al. 2013) and others decreased (Jayakumar et al. 2005; Chua et al. 2007; Venkatasubramanian, 2010) GM volume in this region. A possible explanation for these inconsistencies, is that most studies have used small sample sizes and therefore may have been under-powered. For example, in the most recent meta-analyses (Radua et al. 2012; Gao et al. 2017; Shah et al. 2017), out of a total of 37 studies included (after accounting for overlapping studies across meta-analyses), 20 had a total sample size of 60 or less. Studies with small sample sizes are likely to result in overestimates of effect size and low reproducibility due to low statistical power (Button et al. 2013); which suggests that several of these individual small studies may have had an increased risk of reporting false positives. In addition to being under-powered, different studies have also varied significantly in terms of their methods such as recruitment criteria, imaging acquisition parameters, preprocessing and statistical analysis (Radua et al. 2012). Furthermore, the vast majority of studies have examined participants from a single research site, raising the possibility that the results might be specific to the characteristics of the local sample investigated.

To overcome some of these limitations, the ENIGMA consortium developed a standardized pipeline detailing data preprocessing and analysis procedures; once data is analysed, single-site results are pooled and summarized in a meta-analysis. This approach has led to unprecedented sample sizes in schizophrenia research, with two recent studies of cortical abnormalities in 4474 patients and 5098 controls (van Erp et al. 2018), and subcortical changes in a smaller, albeit still impressive, sample of 2028 patients and 2540 controls (van Erp et al. 2016). However, although

this approach mitigates some of the main limitations of traditional meta-analysis by reducing the heterogeneity of the pooled single-studies, findings still rely on reported results from individual studies, which may result in limited accuracy (Shah et al. 2017). Multi-centre mega-analyses, involving the preprocessing and integration of data from independent studies in one single statistical analysis, provide an opportunity to overcome this limitation. Gupta et al. (2015) analysed neuroanatomical abnormalities in the first mega-analysis in schizophrenia in a sample comprised of 784 individuals with established schizophrenia and 936 healthy controls collected from 23 sites. Similar mega-analytic efforts focused on the initial stages of the illness, when the effects of confounders are minimal, are still non-existent and evidence is still reliant on small to modest sized studies (X. Gao et al., 2018; Shah et al., 2017).

In light of the limitations of the existing literature, the aim of this study was to use a multi-centre mega-analytic approach to test for neuroanatomical changes in FEP that are consistent across independent samples. Based on the findings of the recent meta-analyses (Radua et al. 2012; Gao et al. 2017; Shah et al. 2017), we hypothesize that i) patients would show grey matter volume reductions in a distributed bilateral network including fronto-temporal regions and well as the insula and cingulate; ii) grey matter volume in the FEP group would be negatively correlated with severity of symptoms; and iii) given previous reports of progressive neuroanatomical changes in psychosis (Olabi et al. 2011; Vita et al. 2012), grey matter volume in the FEP group would also be negatively correlated with duration of illness.

3.2. Methods

3.2.1. Participants

A total of 1074 participants recruited as part as five Independent studies (Chengdu, China (Gong et al. 2015), London, England (Di Forti et al. 2009) Santander, Spain (Pelayo-Terán et al. 2008) and Utrecht, The Netherlands (Korver et al. 2012)) were included in the analysis. All patients were experiencing their first psychotic episode, defined as the first manifestation of psychotic symptoms meeting criteria for a psychotic disorder, as specified by the DSM-IV (APA, 2000) or ICD-10 (Organization World Health, 1992). Recruitment details are reported in Chapter 2, sections 2.1. Demographic and clinical data for patients and healthy controls within each site are

summarized in Table 3.1.

3.2.2. MRI data acquisition

At all 5 sites, volumetric MRIs were acquired using a T1-weighted protocol. At four sites, the scanner field strength was 3T, and at 2 sites it was 1.5T. The details of the image acquisition sequence are reported in Chapter 2, section 2.2.2.

3.2.3. Data analysis

3.2.3.1. Socio-demographic and clinical parameters

Differences between FEP and HC in sex, age and TIV were assessed with a chi-square and independent-sample t-test for categorical and continuous data respectively, using SPSS v24.

3.2.3.2. Preprocessing

Differences in GM volume between HC and FEP were examined using VBM, as implemented in SPM12 software (http://wwwfilionuclacuk/spm) running under MATLAB 92 (The MathWorks, Inc, Natick, Massachusetts) (Ashburner & Friston, 2005). The following steps were followed for the preprocessing of each site: (1) checking for scanner artefacts and gross anatomical abnormalities for each subject, (2) setting the image origin to the anterior commissure and reorienting the image along the AC-PC line and (3) segmenting the image into grey matter, white matter and CSF maps. Next, all available images were used to create a study-specific template as implemented by the DARTEL toolbox (Ashburner, 2007). This procedure warps the GM and WM partitions into a new study-specific reference space representing an average of all the subjects included in the analysis, thus maximizing accuracy and sensitivity (Yassa & Stark, 2009). Finally, GM volume maps were normalized to the Montreal Neurological Institute (MNI) template and subsequently smoothed with an 8mm Gaussian filter. A “modulation step” was also included in the normalization step to preserve the information about the absolute GM values (Mechelli et al. 2005). The final smoothed, modulated, normalized data were used for the statistical analysis.

3.2.3.3. Statistical analysis

Statistical analysis was carried out using an analysis of variance (ANCOVA), with diagnostic group

and scanning site as factors, resulting in 10 experimental groups. Age and gender were included as covariates of no interest. The option of proportional scaling was selected to remove confounding driven by global differences. Neuroanatomical alterations in patients with FEP relative to HC consistent across the five datasets were identified using the "inclusive masking" option as implemented in SPM software. This option allowed us to test for voxels which showed (i) an overall statistically significant difference between patients and healthy controls across all sites (p<0.05 FWE corrected) and (ii) at least a strong trend at each site (p<0.05 uncorrected). Specifically, this consisted on the following steps in SPM: i) comparing all FEP against all HC at p<0.05 FWE corrected using an overall main contrast - $FEP_{all\ sites}$ vs $HC_{all\ sites}$ (e.g. $FEP_{all\ sites} < HC_{all\ sites}$), ii) overlaying this contrast with a second set of five FEP vs HC contrasts, one for each site (e.g. $FEP_{site\ 1} < HC_{site\ 1}$) at p<0.05 uncorrected each, and finally iii) identifying voxels of increased/decreased GMV in FEP relative to HC that survived both the overall and the site-level contrasts (Figure 1). This procedure ensured that any overall statistically significant difference across the five sites would also be present at each site, at least at trend level. Statistical inferences were made using a minimum extent threshold of 50 voxels.

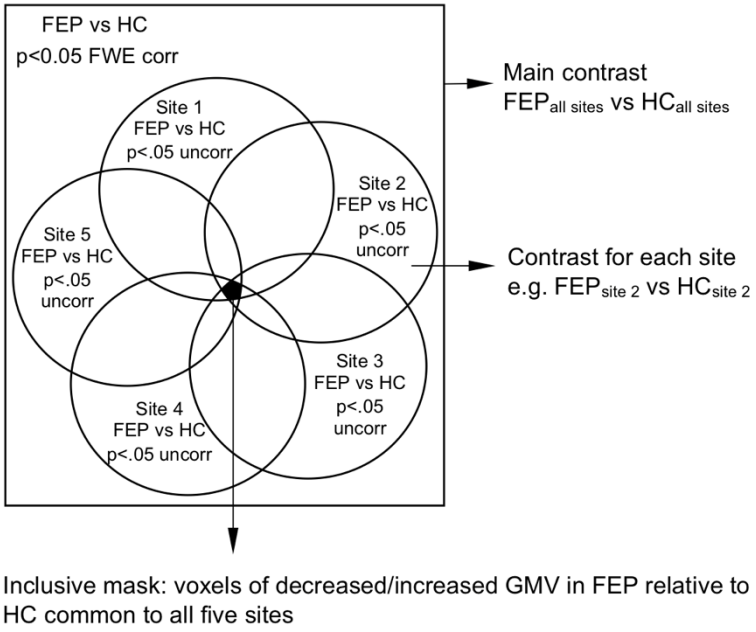


Figure 3.1. Inclusive masking procedure used to identify neuroanatomical abnormalities in FEP relative to HC consistent across all five sites. Left: an overall contrast with all FEP against all HC (p<0.05 FWE corrected) was combined with five site-level contrasts (p<0.05 uncorrected); this allowed us to identify only the voxels that survived both types of contrasts (intersection of all contrasts in black).

The total intracranial volume (TIV) for each image was estimated by first calculating the volume of gray matter, white matter and CSF separately at each voxel from the segmented images; the total volume for each type of tissue was then calculated by summing the respective voxel-level volumes; finally, TIV was obtained by adding the volume of all three tissue types. The effects of symptom severity, illness duration and anti-psychotic medication on the identified clusters were estimated using Pearson's correlation between the values of GMV for the peak coordinate of each statistically significant cluster and each one of the clinical variables of interest. The raw psychotic symptom severity scores (acquired with either PANSS or SANS/SAPS) were first normalized to ensure comparability across sites. This normalisation was achieved using the following formula:

$$\text{New score} = \frac{\text{Individual raw score} - \text{Minimum}}{\text{Maximum} - \text{Minimum}} \quad (3.1)$$

where Minimum and Maximum refer to the lowest and highest score allowed for either PANSS or SAPS/SANS. The resulting disease severity scores were scaled between 0 and 1. Across all sites (except site 1, where all patients were AP-naïve), AP medication dose was estimated by calculating the chlorpromazine equivalent (mg/day) for each individual according to Gardner et al. (2010). Both chlorpromazine equivalent and duration of illness were log transformed. The statistical significance of Pearson's correlation was assessed using a p-value<0.05 with Bonferroni correction for multiple comparisons.

3.3. Results

3.3.1. Socio-demographic and clinical parameters

There were no significant differences between FEP and HC in sex, age and TIV, both when considering all sites together and within each single site. Patients reported comparable median duration of illness across sites (Table 3.1).

3.3.2.1. Decreased GM volume in FEP compared to HC

Relative to HC, FEP showed a widespread pattern of GM volume reduction in fronto-temporal, insular and occipital regions bilaterally (Table 3.2). The largest GM volume reduction was found

in the left gyrus rectus, located in the inferior frontal lobe (Figure 3.1).

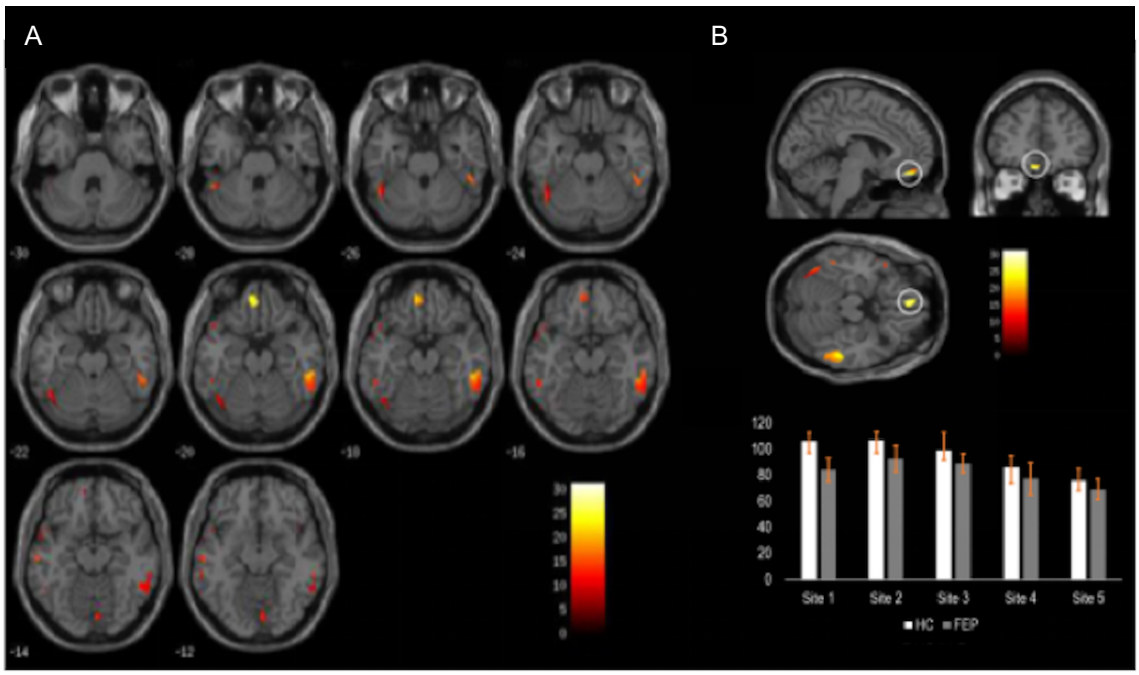


Figure 3.2. GM volume decreases in FEP relative to HC. **A.** Regions showing statistically significant decreases in FEP relative to HC across the whole brain. **B.** Location of the gyrus rectus (straight gyrus) where the largest GM volume decrease was found and mean and standard deviation of the GM volume in this region for each site.

Negative correlations were found between GM volume in this region and severity of both positive and negative symptoms as well as duration of illness (Table 3.4). The left lingual and inferior temporal gyri also showed statistically significant negative correlations with both positive and negative symptoms as well as with duration of illness (Table 3.4).

Table 3.1. Demographic and clinical characteristics for FEP and HC for each site and total sample.

		Chengdu, China (N=240)		London, England (N=168)		Santander A, Spain (N=257)		Santander B, Spain (N=223)		Utrecht, The Netherlands (N=186)		Combined data (N=1074)	
		HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP
N		118	122	92	76	113	144	78	145	101	85	502	572
Sex (%)	M	56 (48)	55 (45)	37 (40)	41(54)	70 (62)	81 (61)	48 (61)	89 (61)	69 (68)	68(80)	280 (56)	341 (60)
	F	62 (52)	67 (55)	55 (60)	35 (46)	43 (38)	56 (39)	30 (39)	56 (39)	32 (32)	17 (20)	222 (44)	231 (40)
		χ^2= ns		χ^2= ns		χ^2= ns		χ^2= ns		χ^2= ns		χ^2= ns	
Age M(SD)		25.8 (8.0)	27.0 (7.3)	26.5 (6.5)	27.0 (6.8)	29.7 (7.7)	29.3 (8.1)	28.0 (7.4)	29.5 (8.7)	26.8 (8.2)	25.4 (5.9)	27.8 (7.5)	27.7 (8.0)
		t= ns		*t*= ns		*t*= ns		*t*= ns		*t*= ns		*t*= ns	
TIV (L) M(SD)		1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.4 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)
		t= ns		*t*= ns		*t*= ns		*t*= ns		*t*= ns		*t*= ns	
Positive symptoms M(SD)		-	24.5 (6.9)[a]	-	13.7 (5.5)[a]	-	14.3 (4.4)[b]	-	13.5 (4.3)[b]	-	15.8 (6.3)[a]	-	-
Negative symptoms M(SD)		-	18.6 (8.6)[a]	-	15.7 (6.0)[a]	-	6.2 (5.0)[c]	-	6.2 (5.0)[d]	-	16.2 (6.9)[a]	-	-
Duration of illness (years) Med (IQR)		-	0.3 (0.9)	-	1.1 (1.3)	-	0.4 (0.7)	-	0.3 (1.0)	-	0.6 (1.4)	-	-

TIV: total intra-cranial volume; L: liters; M: male; F: female; FEP: first episode psychosis, HC: healthy controls; [a]PANSS: Positive and Negative Symptoms Scale; [b]SAPS: Scale for the Assessment of Negative Symptoms; [c]SANS: Scale for the Assessment of Negative Symptoms; ns: $p>0.05$.

Table 3.2. Brain regions of decreased GM volume in FEP relative to the HC.

Region	Peak MNI Coordinates (x,y,z)	Cluster Size (No of voxels)	z	p
L gyrus rectus (straight gyrus)	-6,34,-21	159	9.6	.002
L medial orbital gyrus	-9,54,-15		8.4	
L superior temporal pole	-21,8,-32	119	9.3	.004
L fusiform gyrus	-20,2,-42		8.8	
R inferior temporal gyrus	56,-36,-20	607	9.1	<.001
R middle temporal gyrus	62,-32,-12		8.7	
L inferior temporal gyrus	-51,-52,-27	239	8.9	.001
R fusiform gyrus	-46,-58,-22		8.6	
L middle temporal gyrus	-58,-21,-14	161	8.8	.002
R lingual gyrus	2,-80,-10	106	8.7	.004
L middle temporal gyrus	-52,-42,-18	63	8.5	.009
L superior temporal gyrus	-48, 18,-16	86	8.4	.006
R insula	45,18,-8	88	8.3	.006

L: left; R: right.

3.3.2.2. Increased GM volume in FEP compared to HC

A significant increase in GM volume in FEP relative to control was found in the right superior temporal gyrus (MNI coordinates: 38,16,-38; cluster size: 338; z=81; *p*<001) (Figure 3.2). The volume of this region was not significantly associated with severity of positive (*r*=.05, *p*=.785) or negative (*r*=.07, *p*=.801) symptoms, duration of illness (*r*=.02, *p*=.967) and anti-psychotic medication (*r*=-.08, *p*=.794).

Table 3.3. Pearson's correlations between regions showing GM volume changes in FEP relative to the HC and symptom severity, illness duration and anti-psychotic medication.

	Positive symptoms	Negative symptoms	Duration of illness	Anti-psychotic medication
Decreased GVM in FEP relative to HC				
L gyrus rectus	**-.31**	**-.20**	-.10	-.08
L med. orbital gyrus	**-.17**	**-.17**	-.03	-.06
L sup. temporal pole	-.06	-.07	-.09	-.04
L fusiform gyrus	-.05	-.05	-.11	.02
R inf. temporal gyrus	-.11	-.04	-.08	-.04
R mid. temporal gyrus	.04	.07	-.01	-.02
L inf. temporal gyrus	**-.17**	-.13	-.10	-.11
R fusiform gyrus	**-.15**	-.12	-.08	-.09
L mid. temporal gyrus	.09	.08	.06	.02
R lingual gyrus	**-.20**	**-.16**	**-.18**	-.13
L mid. temporal gyrus	-.07	-.06	-.11	-.08
L sup. temporal gyrus	.02	-.04	-.06	-.03
R insula	.03	-.02	**-.15**	-.08
Increased GVM in FEP relative to HC				
R sup. temporal gyrus	.05	.07	.02	-.08

L: left; R: right; med: median; sup: superior; inf: inferior; mid: middle. Statistical inferences were made at $p<0.05$ after Bonferroni correction for multiple comparisons based on the number of regions; this resulted in a *p*-value of 0.05/14 = 0.0035. Statistically significant correlations are shown in bold.

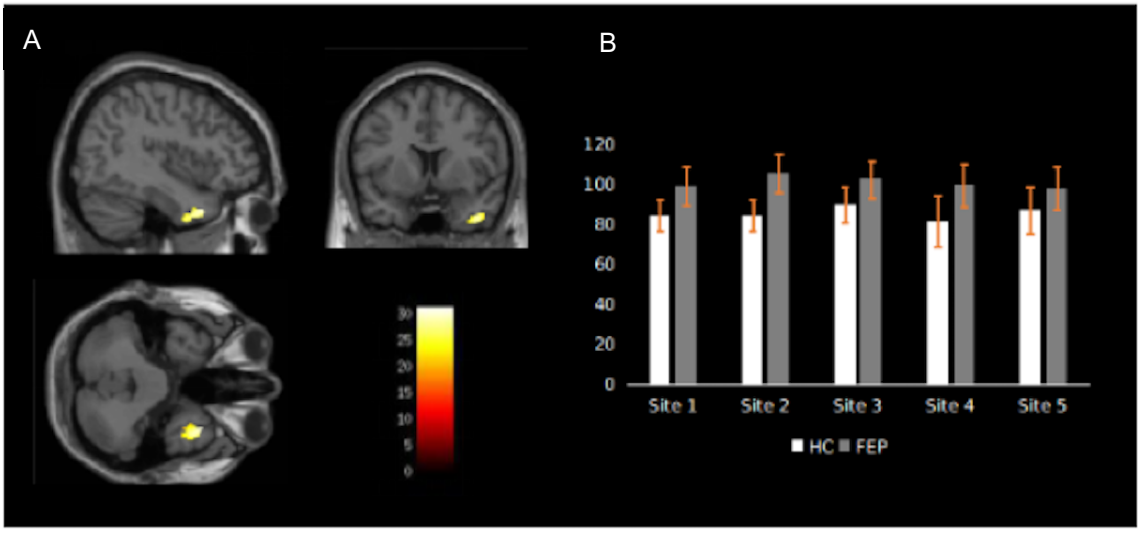


Figure 3.3. GM volume increases in FEP relative to HC. **A.** Location of the right superior temporal gyrus

where the GM volume increase in FEP relative to HC was found. **B.** Mean and standard deviation of the GM volume in this region for each site.

3.4. Discussion

Most previous studies on the neuroanatomical basis of FEP have used small samples recruited within a single site, and have yielded heterogeneous findings (Radua et al. 2012; Gao et al. 2017; Shah et al. 2017). The aim of this study was to use a multi-centre mega-analytic approach to identify neuroanatomical changes in FEP that are expressed consistently across several independent studies. As hypothesized, we found a widespread bilateral pattern of GM volume reductions in fronto-temporal, insular and occipital regions. Some of these effects, particularly in the gyrus rectus and the lingual gyrus, were correlated with symptom severity and duration of illness. In addition, an increase in GM volume was found in the right superior temporal lobe. Critically, all patients were experiencing their first episode of psychosis, one of the five samples was medication-naïve and an additional two were medication-naïve or with limited (up to 6 weeks) lifetime exposure to antipsychotics; this means the current results are unlikely to be explained by illness chronicity and medication effects. In what follows, we discuss the brain structures that emerged from this study as well as their main role in the psychopathology of the early stages of psychosis.

Orbifrontal cortex

A significant GM volume reduction was found in two sub-regions of the orbifrontal cortex (OFC), namely the gyrus rectus (straight gyrus) and the orbital gyrus (Buchanan et al., 2004; Nakamura et al., 2007). Grey matter deficits in the OFC have been reported in established psychosis (e.g. Kim et al. 2017; Kong et al. 2015; Rimol et al. 2012; Xu et al. 2017) and, to a lesser extent, in FEP (e.g. Huang et al. 2015; Keymer-Gausset et al. 2018; Liao et al. 2015), in keeping with the “hypofrontality” hypothesis of psychosis; although increases in this region have also been observed (Gao et al. 2017). The OFC has been implicated in multiple functions, including cognitive flexibility, reward learning and decision making (see Kringelbach 2005 and Schoenbaum et al. 2009 for a review), most of which are impaired in people with psychosis (Murray et al. 2008; Aas et al. 2014; Strauss et al. 2014; Premkumar et al. 2015). The gyrus rectus was the region

with the most pronounced GM volume reduction within the OFC and the whole brain. Consistent with our finding, this region has been reported to be decreased in FEP regardless of antipsychotic medication status in a recent meta-analysis (Shah et al. 2017). GM volume reduction in this region was also found in the largest single-site VBM study of first-episode patients to date which included 93 FEP participants and 175 controls (Meisenzahl et al. 2008); although evidence for normal volume has also been reported (Roiz-Santiáñez et al. 2011; Takayanagi et al. 2011). As hypothesized, GM volume in the gyrus rectus was inversely related to positive and negative symptoms – consistent with previous studies (Szendi et al. 2006; Sans-Sansa et al. 2013; Kim et al. 2017) – and negatively correlated with duration of illness, once again consistent with previous studies (Sapara et al., 2007).

Insula

Despite inconsistences across individual studies, most of the existing literature indicates deficits in the insular cortex of people with FEP, albeit with some inconsistencies in the exact location of the effect (Nekovarova et al. 2014; Gao et al. 2017; Shah et al. 2017; ONeill et al. 2018). Here it was the anterior part of the insula that showed reduced GM volume. This region plays an important role in salience processing (Menon & Uddin, 2010), emotional appraisal and social cognition (Eckert et al. 2009), all of which are affected in psychosis (Wylie & Tregellas, 2010). Notably, grey matter deficits in the insula, as well as in the gyrus rectus and superior temporal gyrus, have also been found in individuals at ultra-high risk for psychosis who later transition to psychosis (Smieskova et al. 2010); this suggests this region may represent a neuroanatomical signature of vulnerability to psychosis. Furthermore, GM volume reductions in this region have been shown to be above and beyond ethnic variations in incidence and clinical expression (Gong et al. 2015).

Temporal cortex

Reductions in temporal regions are amongst the most replicated findings in psychosis, including in FEP (Chan et al., 2011; Radua et al., 2012; Shah et al., 2017). In this study, several temporal regions showed GM volume deficits, namely the superior, middle and inferior gyri as well as the temporal portion of the fusiform gyrus bilaterally. GM volume deficits in the left superior temporal gyrus are thought to play a central role in auditory verbal hallucinations in FEP patients (Benetti

et al., 2015; Modinos et al., 2013), possibly due the role of this region in language perception and processing; it has been suggested that impairment to this region may lead to a misattribution of internal speech (Frith & Done 1988; Mechelli et al. 2007). The fusiform gyrus is also thought to play an important role in the psychopathology of psychosis, mainly due to its contribution to facial recognition (Haxby, Hoffman, & Gobbini, 2002; Haxby et al., 2000), which is impaired in psychosis (see Green et al. 2015 and Barkl et al. 2014 for a review) and is often seen as a proxy for the social cognition deficits characteristic of the illness (Green et al., 2015). Perhaps more challenging to interpret is the significant increase in GM volume in right superior temporal gyrus. Nevertheless, increases in patients relative to controls across the brain, including the temporal cortex, have been reported before (J.-J. Kim et al., 2003; J. S. Lee et al., 2011; Radewicz, Garey, Gentleman, & Reynolds, 2000; Taylor et al., 2005), and are typically interpreted in terms of a "compensatory mechanism" (Guo, Palaniyappan, Liddle, & Feng, 2016) or a transient inflammation resulting from increased apoptotic activity, i.e. removal of cells that have been programmed to die (Berger et al. 2003; Adler et al. 2005).

Lingual gyrus

Evidence supporting structural abnormalities in the lingual gyrus in FEP has not been as consistent, with some studies reporting decreased (Ellison-Wright et al. 2008) and others increased (Gao et al. 2017) GM volume. Such inconsistency may be explained by medication status, as shown by Shah et al. (2017), where GM volume of the lingual gyrus was decreased in antipsychotic naive FEP patients but increased in FEP patients under-going antipsychotic treatment. However, in our study, which included both samples with and without exposure to antipsychotics, there was a consistent reduction in the lingual gyrus in the five sites, suggesting that reductions in this region may be present above and beyond medication status. The lingual gyrus is involved mainly in visual processing (Lee et al. 2000; Hahn et al. 2006) which are well documented in psychosis (see Butler et al. 2008 and Silverstein & Keane 2011 for a review) and are also thought to underlie some of the cognitive impairments characteristic of the illness (Surti et al. 2011; Surti & Wexler, 2012; Contreras et al. 2018). The lingual gyrus also contributes to the evaluation of emotional faces (Fusar-Poli et al. 2009) which, together with the deficits found in the fusiform gyrus, may explain social cognition impairments in psychosis (Green et al., 2015).

Limitations

A first limitation of this study was that clinical data was acquired using different instruments (positive symptoms were assessed with either the PANSS or SAPS and negative symptoms with the PANSS or SANS). We overcame this limitation by normalizing individual scores within each scale as in previously studies (Gong et al., 2018). The resulting scores were highly correlated (r=.87) with automated methods to convert scores between these two widely used scales (van Erp et al. 2014). A further limitation is that, while most FEP participants had limited or no exposure to antipsychotics, participants from sites 4 and 5 were medicated.

3.5. Conclusion

This study aimed to overcome the limitations of small and single-site studies by conducting a multi-centre mega-analysis of neuroanatomical abnormalities in FEP. To the best of our knowledge this is the largest VBM study in FEP to date. We found a widespread pattern of fronto-temporal, insular and occipital GM volume reductions in FEP that were expressed consistently across five independent studies; this provides evidence for reliable neuroanatomical alterations in FEP, expressed above and beyond site-related differences in recruitment criteria and scanning parameters. With the increasingly availability of larger datasets, future multi-centre mega-analyses could investigate the diagnostic specificity of these findings by integrating data collected from people with different psychiatric diagnoses (Ellison-Wright & Bullmore, 2010).

Chapter 4

Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications

This chapter is based on the paper entitled Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications published in Neuroscience and Biobehavioural Reviews.

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58-75.

4.1. Introduction

In the last two decades, neuroimaging studies of psychiatric and neurological patients have relied on mass-univariate analytical techniques (e.g. statistical parametric mapping). These studies typically compared patients with a diagnosis of interest against disease-free individuals and reported neuroanatomical or neurofunctional differences at group-level. The simplicity and interpretability of this approach have led to significant advances in our understanding of the neurobiology of psychiatric and neurological disorders. Mass-univariate analytical techniques, however, suffer from at least two significant limitations. First, statistical inferences are drawn from multiple independent comparisons (i.e. one for each voxel) based on the assumption that different brain regions act independently. This assumption, however, is not in line with our current understanding of brain function in health and disease (Biswal et al., 2010; Fox et al., 2005); for example, several psychiatric and neurological symptoms are best explained by network-level changes in structure and function rather than focal alternations (Kennedy & Courchesne, 2008; Mulders, van Eijndhoven, Schene, Beckmann, & Tendolkar, 2015). Second, mass-univariate techniques can be used to detect differences between groups but do not allow statistical inferences at the level of the individual. In contrast, a clinician has to make diagnostic and treatment decisions about the person in front of them. These two limitations may have contributed to the limited translational impact of neuroimaging findings in everyday clinical practice so far.

In an attempt to overcome these limitations, the neuroimaging community has developed a growing interest in machine learning, an area of artificial intelligence that aims to develop algorithms that discover trends and patterns in existing data and use this information to make predictions on new data. This is achieved through the use of computational statistics and mathematical optimization (Hastie, 2009). Machine learning methods are multivariate and therefore take the inter-correlation between voxels into account, thereby overcoming the first limitation of mass-univariate analytical techniques. In addition, machine learning methods allow statistical inferences at single subject level and therefore could be used to inform diagnostic and prognostic decisions of individual patients, thereby overcoming the second limitation of mass-univariate analytical techniques (Arbabshirani et al., 2017). Machine learning methods can be divided into two broad categories: supervised and unsupervised learning. In supervised machine

learning, one seeks to develop a function which maps two or more sets of observations to predefined categories or values. In contrast, unsupervised methods seek to determine how the data is organised without using any a priori information supplied by the operator; here the main objective is to discover unknown structure in the data (Hastie, 2009).

Over the past decade, several machine learning methods have been applied to neuroimaging data from psychiatric and neurological patients with varying degrees of success (Arbabshirani et al., 2017; Wolfers et al., 2015). The most popular amongst these methods is Support Vector Machine (SVM), a supervised technique that works by estimating an optimal hyperplane that best separates two classes. When these classes are not linearly separable, SVM uses external functions (kernels) that map the original data into a new feature space where the data become linearly separable (Pereira & Mitchell, 2008; Vapnik, 1995). The application of SVM (and most traditional machine learning algorithms in general) typically involves two steps prior to classification: “feature extraction” and “feature selection”. Feature extraction involves the transformation of the original data into a set of “features” that can be used as input. This may consist of transforming each three-dimensional image into a column vector of features where each value corresponds to the intensity (e.g., brain activity or grey matter volume) of a single voxel. Feature selection, on the other hand, involves the selection of a subset of the original features. The aim is to discard any features considered to be either of minimal importance or redundant for the task at hand. In neuroimaging, this may consist of manually selecting regions of interest or, alternatively, use data-driven approaches such as recursive feature elimination. Whilst feature selection represents an optional step, the use of feature extraction is typically a prerequisite of SVM. Indeed, despite its popularity, SVM has been criticised for not performing well on raw data and thus requiring the significant expertise to extract informative features from the data, which are then used for classification. While SVM remains a very popular technique within the neuroimaging community, an alternative family of machine learning methods known as deep learning (Bengio, 2009) is gaining considerable attention in the wider scientific community (Arbabshirani et al., 2017; Calhoun & Sui, 2016; LeCun et al., 2015). Deep learning methods are a type of representation-learning methods, which means that they can automatically identify the optimal representation from the raw data without requiring prior feature extraction (LeCun et al.,

2015). This is achieved through the use of a hierarchical structure with different levels of complexity, which involves the application of consecutive nonlinear transformations to the raw data. These transformations result in increasingly higher levels of abstraction, where higher-level features are more invariant to the noise present in the input data than lower level ones (LeCun et al., 2015). Inspired by how the human brain processes information, the building blocks of deep learning neural networks – known as “artificial neurons” – are loosely modelled after biological neurons. Artificial neurons are organized in layers. A deep neural network consists of an input layer, two or more hidden layers and an output layer. The input layer comprises the data inputted into the model (e.g. voxel intensity); the hidden layers learn and store increasingly more abstract features of the data; these features are then fed to the output layer that assigns the observations to classes (e.g. controls vs. patients). Learning is achieved through an iterative process of adjustment of the interconnections between the artificial neurons within the network, much like in the human brain (Bengio, 2009). An essential aspect of deep learning that differentiates it from other machine learning methods is that the features are not manually engineered; instead, they are learned from the data, resulting in a more objective and less bias-prone process. Besides, the ability to achieve higher orders of abstraction and complexity relative to other machine learning methods such as SVM makes deep learning better suited for detecting complex, scattered and subtle patterns in the data (Plis et al., 2014).

From a historical perspective, the use of deep learning in scientific research can be traced back to the perceptron (i.e. the original version of the artificial neuron), which many researchers refer to as the first machine learning algorithm (McCulloch & Pitts, 1943). After several setbacks, the pioneering work of Warren McCulloch and Walter Pitts resulted in the development of what is now known as artificial neural networks. However, such networks were able to handle a limited number of hidden layers. It was only in the 2000s that researchers developed a new approach for training artificial neural networks that allowed the inclusion of several hidden layers resulting in greater levels of complexity (Hinton et al., 2006). This breakthrough led to the development of a new family of machine learning methods - known as deep learning - which has been shown to outperform previous state-of-the-art classification methods in areas such as speech recognition, computer vision and natural language processing (Krizhevsky et al., 2012; Le, 2013).

The use of deep learning could be particularly useful in the investigation of psychiatric and neurological disorders, which tend to be associated with subtle and diffuse neuroanatomical and neurofunctional abnormalities. Since high-level features can be more robust against noise in the input data, deep architectures may be more suitable to identify diagnostic and prognostic biomarkers than conventional machine learning methods. Deep learning techniques might also provide an ideal tool to investigate the multi-faceted nature of psychiatric and neurological disorders since cross-modality relationships (e.g. neuroimaging and genetics) are likely to occur at an even deeper level (Plis et al., 2014). In addition to these conceptual differences, the use of deep learning to investigate psychiatric and neurological disorders has the practical advantage of not requiring manual feature extraction (LeCun et al., 2015). Therefore, it is unsurprising that an increasing number of neuroimaging studies are using deep learning to elucidate the neural correlates of these disorders (J. Kim et al., 2016; Payan & Montana, 2015; Plis et al., 2014).

Given the insurgence of interest in deep learning within the field of neuroimaging, this review aims to give a brief overview of deep learning and potential applications to the investigation of brain-based disorders. In the first part of the review, we outline the underlying concepts of deep learning. To achieve this, we will use one of the simplest deep learning structures, i.e. the multilayer perceptron, to illustrate the steps of training and testing. This will be followed by a brief description of the most common deep learning architectures used in the field of neuroimaging, including stacked autoencoders, deep belief networks and convolutional neural networks. The second part of this article aims to summarise the studies that have applied deep learning to neuroimaging data to investigate psychiatric and neurological disorders. Finally, in the third part of the review, we discuss the main themes that have emerged from our review of the existing literature, and make a number of suggestions for future research directions.

4.2. Overview

Deep learning refers to the training and testing of multi-layered neural networks that are capable of learning complex structures and achieve high levels of abstraction. There are two main types of deep learning models which differ with respect to how the information is propagated through

the network. In feedforward networks, the information is propagated through the network in just one direction, from the input to the output layer. Recurrent networks, in contrast, contain feedback connections that allow the information from past inputs to affect the current output. These connections enable the information to persist within the neural network, akin to a form of memory, and this allows the models to process sequential data, such as speech and language, in a natural way.

The implementation of deep learning in the context of supervised classification problems involves two main steps. In the first step, the so-called *training phase*, a subset of the available data known as the *training set* is used to optimize the network's parameters to perform the desired task (classification). In the second step, the so-called *testing phase,* the remainder subset which is known as the *test set* is used to assess whether the trained model can blind-predict the class of new observations. When the amount of available data is limited, it is also possible to run the training and testing phases several times on different training and test splits of the original data and then estimate the average performance of the model – an approach known as cross-validation. The two phases of training and testing are not a specific feature of deep learning but are used in conventional machine learning methods.

In this section, we will discuss the use of feedforward deep learning for classification problems. We will start with the multilayer perceptron (MLP), the simplest deep learning architecture, to illustrate three important aspects of deep learning – network structure, training and testing. We will then describe more complex networks, including stacked autoencoders and deep belief networks. Finally, we will describe the increasingly popular convolutional neural networks (CNN), an important adaptation of the MLP that has come to be considered the state-of-the-art for computer vision.

4.2.1. Multilayer perceptron

4.2.1.1. Network structure

MLPs are organized in a layer-wise structure where each layer stores increasingly more abstract representations of the data (Figure 4.1). The first layer is the input layer where the data is entered

into the model. In neuroimaging, the data can be represented as a one-dimensional vector with each value corresponding to the intensity of one voxel. The last layer is the output layer which, in the context of classification, yields the probability of a given subject belonging to one group or the other. The layers between the input and output layers are called hidden layers, with the number of hidden layers representing the depth of the network. Each layer comprises a set of artificial neurons or "nodes" (Figure 4.1A) in which each neuron is fully connected to all neurons in the previous layer (Figure 4.1B). Each connection is associated with a weight value, which reflects the strength and direction (excitatory or inhibitory) of each neuron input, much like a synapse between two biological neurons. Compared to other networks, the multilayer perceptron has a generic structure and therefore is a general-purpose network, i.e. they are not built to process a specific type of data. However, due to its fully connected layers and respective high number of weights to estimate, it is commonly applied to non high-dimensional data.

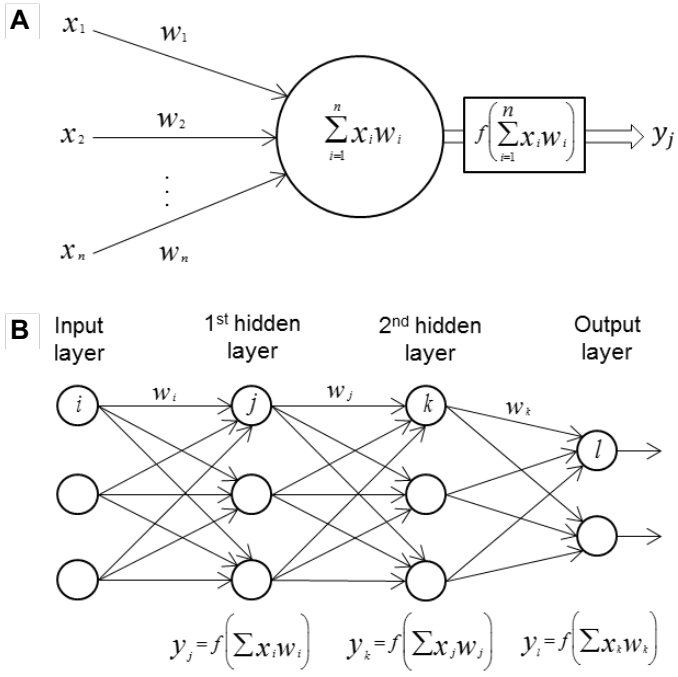


Figure 4.1. Artificial neuron and deep neural network. **A.** The building block of deep neural networks – artificial neuron or node. Each input x_i has an associated weight w_i. The sum of all weighted inputs, $\Sigma x_i w_i$, is then passed through a nonlinear activation function f, to transform the pre-activation level of the neuron to an output y_j. For simplicity, the bias terms have been omitted. The output y_j then serves as input to a node in the next layer. Several activation functions are available, which differ with respect to how they map a pre-

activation level to an output value. The most commonly activation functions used are the rectifier function (where neurons that use it are called rectified linear unit (ReLU)), the hyperbolic tangent function, the sigmoid function and the softmax function. The latter is commonly used in the output layer as it can compute the probability of multiclass labels. **B.** Example of a feedforward multilayer neural network (also referred to as multilayer perceptron) with two classes, in which the nodes in one layer are connected to all neurons in the next layer (fully connected network). For each neuron *j* in the first hidden layer, a nonlinear function is applied to the weighted sum of the inputs. The result of this transformation (y_j) serves as input for the second hidden layer. The information is propagated through the network up to the output layer, where the softmax function yields the probability of a given observation belonging to each class.

Unlike SVM, which relies on expert designed transformations to handle nonlinearly separable classes, the structure of neural networks itself allows the transformation of the input space. The consecutive layers perform a cascade of nonlinear transformations that distort the input space allowing the data to become more easily separable (Figure 4.2). The optimal number of layers and nodes within each layer are not estimated as part of the learning process itself but are defined *a priori* and are called hyperparameters. It should be noted that the development of algorithms to find optimum values of these hyperparameters is an active area of research, and that at present there are no fixed rules (Bergstra, Bardenet, Bengio, & Kégl, 2011; Gelbart, Snoek, & Adams, 2014).

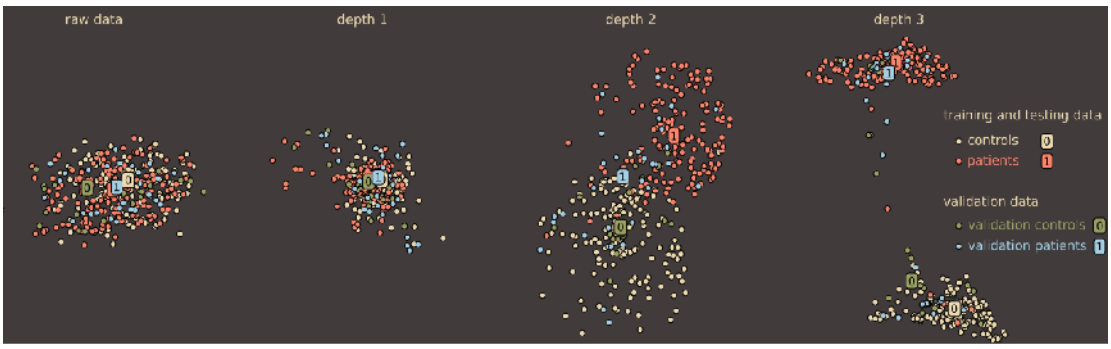


Figure 4.2. Effect of the depth of the model plotted using a neighbourhood-based embedding. With more hidden layers, the data becomes more easily separable due to nonlinear transformations along the network (Plis et al., 2014).

4.2.1.2. Training

Traditionally, neural networks can learn through a gradient descent-based algorithm. The gradient descent algorithm aims to find the values of the network weights that best minimize the error (difference) between the estimated and true outputs. Since MLPs can have several layers, in order to adjust all the weights along the hidden layers, it is necessary to propagate this error backward (from the output to the input layer). This propagation procedure is called backpropagation, and allows the network to estimate how much the weights from the lowers layer need to be changed by the gradient descent algorithm. Initially, when a neural network is trained, the weights are set at random. When the training set is presented to the network, this forward propagates the data through the nonlinear transformation along the layers. The estimated output is then compared to the true output, and the error is propagated from the output towards the input, allowing the gradient descent algorithm to adjust the weights as required. The process continues iteratively until the error has reached its minimum value. The backpropagation algorithm does not work well with the original models of DNNs that were based on sigmoid and hyperbolic tangent nonlinearities. In these models, the information of the error becomes increasingly smaller as it propagates backward from the output to the input layer, to a point where initial layers do not get useful feedback on how to adjust their weights – an issue known as the vanishing gradient problem. Therefore, initially, the use of backpropagation yielded poor solutions for networks with three or more hidden layers (Schmidhuber, 2015). In 2006, however, Hinton and colleagues put forward the idea of “greedy layerwise training”, which consists of two steps: 1) an unsupervised step, where each layer is trained individually and 2) a supervised step, where the previously trained layers are stacked, one additional layer is added to perform the classification (the output layer), and the whole network parameters are fine-tuned (Hinton et al., 2006). This breakthrough led to the fast-growing interest in deep learning and enabled the development of at least two types of pre-trained networks that have shown promising results: stacked autoencoders and deep belief networks. It should be noted that these methods are not actual classifiers themselves; instead, they are networks that are pre-trained to learn useful patterns in the data and then fed to a real classifier at the final layer. These two types of networks and their unique characteristics are described in section 4.2.2 and 4.2.3.

4.2.1.3. Testing

As with traditional machine leanrning models, the performance of a deep neural network can be evaluated by several performance measures, such as sensitivity, specificity, accuracy and F-score. Sensitivity refers to the proportion of true positives correctly identified (e.g. the proportion of subjects that were predicted as patient and are true patients), and specificity refers to true negatives correctly identified (e.g. the proportion of subjects that were predicted as healthy controls and are true healthy controls). The accuracy of a classifier represents the overall proportion of correct classifications. The statistical significance of this overall accuracy can be tested using parametric tests such as permutation testing, which measures how likely the observed accuracy would be obtained by chance. Metrics such as F-score and balanced accuracy, which take into account each group's sample size, are particularly useful in cases where classes are unbalanced. The F-score is the weighted harmonic mean of the test's precision and recall[5]. Balanced accuracy, on the other hand, corresponds to the average accuracy obtained on either class (Brodersen, Ong, Stephan, & Buhmann, 2010).

4.2.1.4. Parameters, hyperparameters, and hyperparameters tuning

Similarly to the coefficients in a linear or logistic regression, the weights and bias are the parameters of a deep learning network. As mentioned in section 4.2.1.2. these are estimated form the data during training via the optimization of a loss function, usually through the use of a gradient-descent based algorithm in combination with backpropagation. Hyperparameters on the other hand, are, broadly speaking, a configuration that is external to the model and whose value cannot be estimated from data. However, some model's aspects (e.g. the step-size in the recursive feature elimination (RFE) procedure) can be considered hyperparameters and it might be possible to estimate them from the data (e.g. using nested cross-validation); however this is not always possible to do, for example due to computational reasons. The number of layers, the number of neurons within each layer, the activation function, the optimizer, the learning rate, and the regularization strategy are only a few examples of a long list of hyperparameters one has to consider when building a multilayer perception (and deep learning models in general, although specific network architectures may have additional ones). This long list results in an endless

[5] F-score $= \mathbf{2} * \frac{precision * recall}{precision + recall}$, where precision $= \frac{True\ positives}{True\ positives + True\ negatives}$ and recall $= \frac{True\ positives}{True\ positvies + False\ negatives}$

number of possible combinations of hyperparameters. Although automated tuning (e.g. grid search, random search or Bayesian optimization), in which the algorithm is instructed to test alternative values and select those that provide the best result, is possible, it can be computationally expensive. However, with the fast-growing availability of graphical processing units, the application of DL is likely to become less expensive and more feasible in the future. Manual optimization, i.e., trial and error, is arguably the most common type of hyperparameter tuning in brain disorders research. Although guidelines are available (Bengio, 2012), manual adjustment is mostly based on the intuition of how the network behaves with different hyperparameters. This approach has the advantage of being much less computationally expensive than automatic tuning; however, it requires a great deal of technical expertise and is potentially prone to subjective bias. Importantly, regardless of the tuning method, any study that aims to be replicable should ideally report both the explored set of hyperparameters and the hyperparameters adopted in the final solution. Without the information on the architecture of the DNN and the explored and adopted hyperparameters, any result would be challenging to replicate and as such should be taken with caution.

4.2.1.5. Risk of overfitting and possible strategies

Due to the use of multiple nonlinear transformations, deep networks are highly complex models that involve the estimation of a very large number of parameters. This can lead to the model learning particular fluctuations in the training data that are irrelevant for the purpose of classification – an issue known as “overfitting”. When this happens, the model will perform very well on the training data but will not be able to replicate its performance on unseen data (Srivastava et al., 2014). The risk of overfitting is particularly high in the context of neuroimaging, where the number of data points (e.g. number of voxels) for a subject is much larger than the total number of subjects, resulting in high-dimensional data (Arbabshirani et al., 2017). However, there are a number of strategies that can be used to minimise the risk of overfitting, collectively known as “regularization”. A first strategy involves the use of weight decays (e.g., L1 and L2 norms) to penalise models with very high weights. It has been observed that extreme (very low or very high) weight values in a machine learning model are symptomatic of the model trying to learn the regularities of the data perfectly (Krogh & Hertz, 1992). By forcing weights to remain low, the

network becomes less dependent on the training data and is able to better generalize to unseen data (Nowlan & Hinton, 1992). A second strategy, known as dropout, consists of temporarily removing a random number of nodes and their respective incoming and outgoing connections from the network during training. This means that the contribution of dropped-out neurons to the activation of downstream neurons is temporally removed on the forward pass and that any weight updates are not applied to these neurons on the backward pass. The aim of dropout is to extract different sets of features that can independently produce a useful output, thereby allowing higher levels of generalizability (Srivastava et al., 2014).

4.2.2. Autoencoders

Autoencoders are a special case of feedforward networks which comprise of two main components. The first component, i.e. the “encoder”, learns to generate a latent representation of the input data, whereas the second component, i.e. the “decoder”, learns to use these learned latent representations to reconstruct the input data as close as possible to the original (Figure 4.3A) (Vincent et al., 2010).

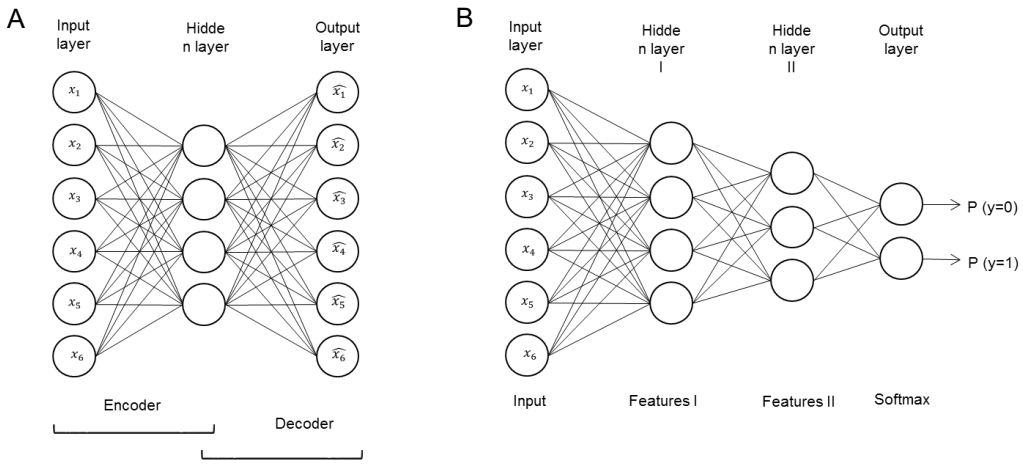


Figure 4.3. Autoencoder. **A.** Shallow or simple autoencoder. In its shallow structure, an autoencoder is comprised of an input layer, that represents the original data (e.g., pixels in an image), one hidden layer that represents the transformed data, and an output layer that reconstructs the original input data. **B.** Stacked autoencoder. Two simple autoencoders are stacked with a 2- class softmax classifier as the final layer. From each simple autoencoder, the output layer is discarded, and the hidden layer is used as the input layer for next autoencoder.

Since an autoencoder does not make use of labels, its training is an unsupervised learning process. In its shallow structure, an autoencoder is comprised of three layers: an input layer, one hidden layer and an output layer. The training to perform the input-copying task can be useful to extract meaningful features of the input data. This automatic feature extraction can be performed using an error function (or loss function) that encourages the model encoder to have specific characteristics, such as sparsity of the representation (sparse autoencoders) and robustness to noise (denoising autoencoders). Since autoencoders are automatic features extractors, they can also be stacked to create a deep structure to increase the level of abstraction of learned features. In this case, the network is pre-trained, i.e. each layer is treated as a shallow autoencoder, generating latent representations of the input data. These latent representations are then used as input for the subsequent layers before the full network is fine-tuned using standard supervised learning (Figure 4.3B) (Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007). In neuroimaging, the most common application of autoencoders has been for pre-training networks (Heinsfeld, Franco, Craddock, Buchweitz, & Meneguzzi, 2018; Kim, Calhoun, Shim, & Lee, 2016), although they can also be used for other purposes such as dimensionality reduction, akin to principal components analysis (Hazlett et al., 2017) or, more recently, to building normative models (Pinaya, Mechelli, & Sato, 2018).

4.2.3. Deep belief networks

Deep belief networks (DBNs), proposed by Hinton et al. (2006), are technically the first deep learning models. Similar to stacked autoencoders, DBNs are comprised of stacked shallow feature extractors, known as restricted Boltzmann machines (RBMs). An RBM is composed by only two layers: a visible layer and a hidden layer. Just like autoencoders, RBMs also aim to learn and extract useful features from the data. However, RBMs differ from autoencoders with regards to their training processes. RBMs can be interpreted as a stochastic neural network. Therefore, instead of using deterministic functions and the reconstruction error (like the autoencoders), the RBM uses the maximum-likelihood estimation to find a stochastic representation of the input in its hidden layer (latent features). To do this, RBMs are usually trained using a gradient descent algorithm, with the likelihood gradient being performed by an approximation algorithm known as

contrastive divergence (Hinton et al., 2006). Here the input data, stored in the visible layer, are propagated to the hidden layer as in a feedforward network, and the resulting sum of the weighted inputs provides a measure of the neuron activation probability. The activation of hidden neurons can be thought of as the network's internal representation of the data, which is then propagated back to the visible layer in an attempt to reconstruct the input data from the network's internal representation. The network, therefore, learns by adjusting the weights based on the discrepancy between the true and reconstructed data. Similarly to autoencoders, RBMs can be stacked to create a deep network, where the hidden layer representation of one RBM serves as input layer for the following RBM, and the network can learn higher-level features from lower-level ones to arrive at an abstract representation of the data. Furthermore, the neural network corresponding to a trained DBN can be augmented by adding an output layer, where units represent the labels corresponding to the input sample. This results in a standard neural network for classification that can be further trained using supervised learning algorithms.

4.2.4. Convolutional neural networks

Convolutional neural networks (CNNs) are a special type of feedforward neural networks that were initially designed to process images, and as such are biologically-inspired by the visual cortex (LeCun et al., 1998). In addition to the input and output layers, CNN can comprise of three types of layers: a convolutional layer, a pooling layer, and a fully-connected layer (Figure 4.4). The convolutional layer is organized in several feature maps. Every neuron in a feature map is connected to a fixed set of neurons in a local region of the previous layer – the *receptive field* – in such a way that the whole image is covered ("local connectivity"). Within the same feature map, the connections between each neuron and the corresponding *receptive field* share the same weights, whereas different feature maps use different sets of weights ("weight sharing").

As a result of this architecture, a feature map can be thought of as a "feature detector" that scans the whole image for the same pattern. This pattern is usually known as the kernel. Kernels in a CNN are learned during the training process, as opposed to in SVM, where they are defined a priori. In a network with several convolutional layers, each layer codes for increasingly more abstract features (e.g. lines → edges → eyes → face). The pooling layer simply reduces the

number of neurons of the previous convolutional layer. The fully-connected layers are similar to the hidden layers from the conventional MLP where the neurons are connected to all neurons from the previous layer. All combined, the properties of CNN (local connectivity, weight sharing and pooling) result in a significant reduction in the number of parameters, which in turn decreases the likelihood of overfitting, and alleviates computational processing. Historically, CNNs have been specifically designed to process images and to this day this is the most commonly type of data used. In principle however, CNNs can be useful when one wants to make the most out of dependencies among features based on spatial distances, i.e. grid-like data. In psychiatric and neurologic neuroimaging, this is applicable to voxel-level MRI data (Payan & Montana, 2015) including functional MRI (Sarraf & Tofighi, 2016) but also electroencephalograms (Acharya, Oh, Hagiwara, Tan, & Adeli, 2018).

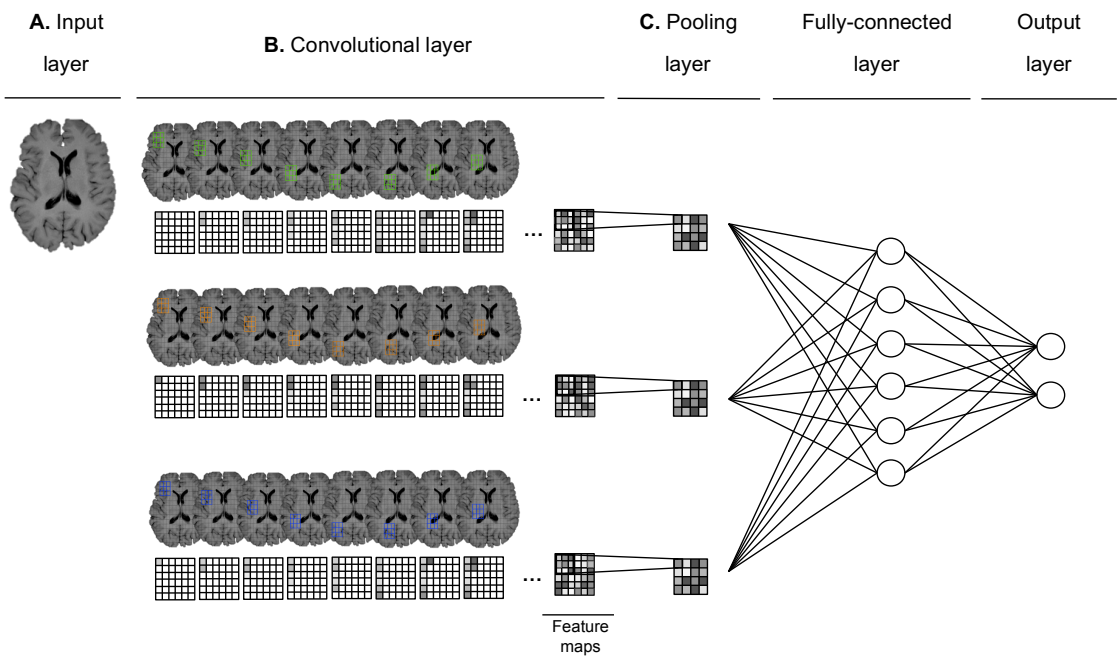


Figure 4.4. Generic structure of a CNN. For illustrative purpose, this example only has one layer of each type; a real-world CNN, however, would have several convolutional and pooling layers (usually interpolated) and one fully-connected layer. **A.** Input layer. In its simplest way, the data is inputted into the network in such a way that each voxel corresponds to one node in the networks. **B.** Convolutional layer. A 3x3 filter or kernel (in green) is used to multiply the spatially corresponding 3x3 nodes in the image. The resulting weighted sum is then passed through a nonlinear function to derive the output value of one node in the feature map. The repetition of this same operation across all possible overlapping receptive fields results in one complete

feature map. The same procedure with different kernels (in orange and blue) will result in separate complete feature maps. **C.** Pooling layer. The size of each feature map can be reduced by taking the maximum value (or average) from a receptive field.

4.3. Review of deep learning studies of psychiatric or neurological disorders

In order to identify previous applications of deep learning in neuroimaging studies of psychiatric or neurological disorders, a search was conducted on 1st August 2016 across several databases (PubMed, IEEE Xplore, Scopus and ArXiv) using the following search terms: ("deep learning" OR "deep architecture" OR "artificial neural network" OR "autoencoder" OR "convolutional neural network" OR "deep belief network") AND (neurology OR neurological OR psychiatry OR psychiatric OR diagnosis OR prediction OR prognosis OR outcome) AND (neuroimaging OR MRI OR “Magnetic Resonance Imaging” OR "fMRI" OR “functional Magnetic Resonance Imaging” OR PET OR "Positron emission tomography"). This review did not include EEG studies, although there is some evidence that deep learning can also be used with this type of data, particularly in epilepsy (Page, Turner, Mohsenin, & Oates, 2014). The initial search yielded a total of 172 articles. As the next step, we screened and cross-referenced these articles for studies that had applied a deep learning model to neuroimaging data to investigate a psychiatric or neurologic condition; this identified a total of 25 articles which were relevant to our review. We organized these articles as follows: i) *diagnostic studies*, which aimed to classify patients from healthy controls, ii) *studies on conversion to illness*, which used baseline scans from individuals identified as being at high risk of developing a psychiatric or neurologic disorder to predict subsequent transition to the illness, and finally iii) *studies predicting treatment response,* which used baseline scans from individuals with a neurological or psychiatric diagnosis to predict subsequent treatment response. These studies are summarised in Tables 1, 2 and 3 which provide the following information: sample size; type of data used as input; whether a whole brain (WB) or region of interest (ROI) approach was used; whether the information inputted into the model comprised of voxel or region-level features; whether feature selection or dimensionality reduction was used before inputting the data into the model; general type of deep learning architecture; diagnostic groups being investigated; and accuracy. Whenever performed, we also report the accuracies obtained for multiclass classifications, which involve discriminating between more than

two classes (e.g. healthy controls vs. mild cognitive impairment vs. Alzheimer's disease).

4.3.1. Diagnostic studies

Studies using deep learning to classify psychiatric or neurological patients from healthy individuals have used a range of neuroimaging modalities including structural MRI (sMRI), resting-state fMRI (rsfMRI), positron emission tomography (PET) and a combination of different modalities (multimodal studies) (see Table 4.1). From Table 1 it can be seen that the vast majority of these studies were carried out in Alzheimer's disease (AD) and its prodromal stage, mild cognitive impairment (MCI). In addition, a smaller number of studies examined psychosis, attention deficit/hyperactivity disorder (ADHD), cerebellar ataxia and temporal lobe epilepsy (TLE). Within each diagnostic category, we first give an overview of the studies that have used a single neuroimaging modality, followed by studies that employed a multimodal approach and, finally, studies that have combined neuroimaging and clinical data within a single classifier.

Mild Cognitive Impairment and Alzheimer Dementia. In one of the first studies using deep learning in AD and MCI, Gupta et al. (2013) argued that, since (i) natural images and brain imaging have similar, and therefore interchangeable, low-level features (e.g. lines and corners) and (ii) natural images, contrary to neuroimaging, are abundant, then natural images could be used to learn low level features which could then be used to identify lesions along the surface and ventricles of the brain. This process, whereby the features learned in one set of data are used to solve a problem in another set of data, is known as "transfer learning". Based on this premise, the authors pre-trained a sparse autoencoder to learn features from natural images, which were then applied to structural MRI data via a CNN, achieving a classification accuracy of 94.7% for AD versus controls, 86.4% for MCI versus controls and 88.1% for AD versus MCI. Consistent with the authors' hypothesis, the method where features were extracted from natural images outperformed the one where the learned features were extracted from the neuroimaging data (93.8%, 83.3% and 86.3% for the same comparisons, respectively). However, a few years later and using a similar approach, Payan and Montana (2015) found comparable classification accuracies using a CNN model with features that were learned from the structural MRI data itself. This could

potentially be explained by the fact that Payan and Montana (2015) used a much larger sample, as well as by the fact that authors used 3D brain images, as opposed to 2D, which possibly contain more useful patterns for classification. Indeed, Payan and Montana (2015) reported that, in general, the models based on 3D outperformed those based on 2D brain images (AD vs. HC (2D/3D)=95.4%/95.4%; AD vs. MCI (2D/3D)=82.2%/86.8%; MCI vs. HC (2D/3D)=90.1%/92.1%). The best accuracy (97.6%) from single modality studies came from Hosseini-Asl et al. (2016), who also used transfer learning. Instead of extracting features from natural images and then fine-tuning the model on Alzheimer's patients and controls, as seen in Gupta et al. (2013), Hosseini-Asl et al. (2016) used one Alzheimer's dataset for pre-training and another independent Alzheimer's dataset to fine-tune the model. By performing the pre-training on an Alzheimer's dataset, this approach allowed for the network to extract generic features related to AD biomarkers, such as the ventricular size, hippocampus shape, and cortical thickness as opposed to more generic low-level features as in Gupta et al. (2013). By using two independent samples during the complete learning process, the final learned features for classification are much less dataset-specific, and should therefore be more generalizable. The final model's architecture was also deeper than in previous studies, which probably also contributed to the high accuracy. Taken collectively, these studies suggest that the application of deep learning to structural MRI data allows the classification of individuals with AD and MCI with high levels of accuracy. Consistent with the increasing popularity of CNN models, studies that have applied either CNN or a combination of AE and CNN have shown better performances compared to those using only AE, although it should be noted that the former group of studies tended to have larger samples than the latter group. In addition, and similar to the trend reported in computer vision competitions and research, the best performances were obtained by the deepest CNN models.

Studies of AD and MCI using resting-state imaging have also achieved promising results. For example, Han et al. (2015) designed a hierarchical convolutional sparse autoencoder (HCSAE), which essentially extracts the most discriminating features from the resting-state data and encodes them in a convolutional manner. This particular arrangement allows for the extraction of the most useful information while conserving abundant detail. The final model classified AD and controls with an 80.0% accuracy and significantly outperformed SVM, which only yielded an

accuracy of 50% (Figure 4.4). While this is a promising result, the model assumed that functional networks were statistic over time – an assumption which underlies the vast majority of machine learning applications to resting-state neuroimaging data. However, recent studies have shown that the network-level functional organization of the brain is dynamic rather than static (Hutchison et al., 2013). Suk et al. (2016) have addressed this issue by developing an approach which classifies people with MCI and healthy controls using a deep autoencoder to extract hierarchical nonlinear relations among 116 brain regions (each region represented the average intensity of the voxels within that region), whilst modelling the inherent functional dynamics of resting-state data. This was also one of the few studies in which the same deep learning model was tested against and surpassed other competing models in two independent datasets (72.6% for dataset 1 and 80.0% for dataset 2), thus providing evidence of replicability, a crucial feature for diagnostic tools. In line with the studies using structural imaging, the best performance for the classification of AD patients with resting-state data was also obtained by a CNN model applied to minimally pre-processed voxel-level data with an accuracy of 96.9% (Sarraf & Tofighi, 2016). These studies provide initial evidence that brain activity at resting state can be useful in identifying MCI and AD patients. We note that, compared to the performances obtained from structural data, deep learning models applied to functional data seem to perform worse. This discrepancy could be explained by the substantial difference in sample size between the two types of studies – while the *smallest* study using structural data included 140 subjects (Hosseini-Asl et al., 2016) the *largest* study using functional data included 62 subjects (H.-I. Suk, Wee, et al., 2016).

With regards to multimodal studies, Liu et al. (2014) applied a stacked autoencoder (SAE) to structural and PET data and successfully distinguished AD and MCI from controls with an accuracy of 87.8% and 76.9%, respectively. Using a very similar dataset, the same team (Liu, Liu, Cai, Che, et al., 2015) achieved a better performance by designing a model where the hidden layers were able to infer the correlations between sMRI and PET, thus better capturing the synergy between the two modalities. This model classified AD and MCI against controls with an accuracy of 91.4% and 82.1%, respectively. Interestingly, the application of the same model to a structural data alone resulted in less impressive accuracies of 82.6% and 72% for AD and MCI, respectively. This discrepancy suggests that the integration of structural and functional data may

improve classification accuracy. However, this conclusion should be drawn with great caution since that the authors did not report classification accuracy for PET data alone.

Finally, four studies have tried combining neuroimaging data with clinical information to build a more robust classification model. For example, Suk and Shen (2013) used a SAE to extract latent features from neuroimaging data (sMRI, PET and CSF), which were then used to predict clinical data (measured using the Mini-Mental State Examination - MMSE - and Alzheimer's Disease Assessment Scale's cognitive subscale - ADAS-cog) and class labels. As the final step, the resulting learned features were used to classify AD and MCI from healthy individuals with an accuracy of 95.9% and 85.0%, respectively. Notably, two more studies (Li et al., 2014; H.-I. Suk, Lee, & Shen, 2015) that have used the same exact sample (taken from the publicly available dataset ADNI; Alzheimer's Disease Neuroimaging Initiative) and the same types of data (sMRI, PET, CSF, MMSE and ADAS-cog) have also reported high accuracies for both AD and MCI despite using different implementations of deep learning. In general, studies combining clinical with neuroimaging data have, in general, reported higher accuracies than studies using single modality or multiple neuroimaging modalities. This is in line with previous studies using conventional machine learning methods (Moradi et al., 2015; Willette et al., 2014) and highlights the usefulness of adding clinical information in the classification of AD and its prodromal phase. It should be noted however, that adding clinical information as predictors may add unwanted circularity to the model. This is because labels are established based on clinical measures and therefore the two are likely to be highly correlated. Thus, it is recommended that the association between these features and the labels is first investigated, and the features with a significant association removed; otherwise the performance is likely to be inflated (Donini et al., 2019).

Table 4.1. Diagnostic studies.

Authors, year	Sample size	Technique	Features	Feature selection/ dimensionality reduction	DL architecture	Comparison	Acc (%)
Gupta et al. (2013)[1]	AD=200	sMRI	WB voxel-level	No	Sparse AE & CNN	HC vs. AD	94.7
	MCI=411					HC vs. MCI	86.4
	HC=232					AD vs. MCI	88.1
						HC vs. AD vs. MCI	85.0
	HC =755	sMRI		No		HC vs. AD	95.4

Payan and Montana (2015)[1]	AD = 755 MCI = 755		WB voxel-level		Sparse AE & CNN	HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	92.1 86.8 89.5
Hosseini-Asl et al. (2016)[1,2]	HC = 70* AD = 70* MCI = 70*	sMRI	WB voxel-level	No	AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	97.6 90.8 95.0 89.1
Chen et al. (2015)[1]	HC = 123 AD = 94 MCI = 121	sMRI	WB voxel-level	Yes	SAE	HC vs. AD HC vs. MCI	89.0 81.7
Liu et al. (2015a)[1]	HC = 204 AD = 180 MCI = 374	sMRI	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	82.6 72.0
Gao and Hui (2016)	HC =117 AD = 51 Lesions = 118	CT	WB voxel-level	No	CNN	HC vs. AD vs. Lesion	87.7
Sarraf and Tofighi (2016)[1]	HC =15 AD = 28	rsfMRI	WB voxel-level	No	CNN	HC vs. AD	96.9
Suk et al. (2016)[1]	HC = 31 MCI =31	rsfMRI	WB region-level	Yes	DAE	HC vs. MCI	72.6
	HC = 25 MCI = 12	rsfMRI	WB region-level	Yes	DAE	HC vs. MCI	81.1
Hu et al. (2016)[1]	HC = 52 MCI = 48	rsfMRI	WB region-level	No	SAE	HC vs. MCI	87.5
Han et al. (2015)[1]	HC = nr AD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. AD	80.0
Liu et al. (2015a)[1]	HC = 77 AD = 85 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	91.4 82.1
Suk et al. (2014)[1]	HC = 101 AD = 93 MCI = 204	sMRI & PET	WB region-level	Yes	DBM	HC vs. AD HC vs. MCI	94.9 80.6
Liu et al. (2014)[1]	HC =77 AD = 65 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	87.8 76.9
Suk et al. (2015b)[1]	HC= 52 AD = 51 MCI = 99	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	HC vs. AD HC vs. MCI HC vs. AD vs. MCI	95.1 80.1 62.9
	HC= 229 AD = 198 MCI = 403	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	HC vs. AD HC vs. MCI HC vs. AD vs. MCI	90.3 70.9 57.7
Liu et al. (2015b)[1]	HC =77 AD = 85 MCI = 169	sMRI & PET & MMSE	WB region-level	Yes	SAE	HC vs. AD HC vs. AD vs. MCI	90.1 59.2
Suk et al. (2015a)[1]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI AD vs. MCI	98.8 90.7 83.7
Li et al. (2014)[1]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	HC vs. AD HC vs. MCI	91.4 77.4
Suk and Shen (2013)[1]	HC = 52 AD = 51	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	HC vs. AD HC vs. MCI	95.9 85.0

	MCI = 99						
Han et al. (2015)[3]	HC = nr ADHD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. ADHD	65.0
Deshpande et al. (2015)[3]	HC =744 ADHD-C = 260 ADHD-I = 173	rsfMRI	WB region-level	Yes	FCC	HC vs. ADHD-C HC vs. ADHD-I ADHD-C vs. ADHD-I	~90.0 ~90.0 95.0
Kuang et al. (2014)[3]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC) ROI (VC) ROI (CC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	37.4 to 71.8** 34.4 to 68.8** 37.1 to 72.7**
Kuang and He (2014)[3]	HC = 42 to 95 ADHD-C = 0 to 77 ADHD-I = 0 to 44 ADHD-H = 0 to 6	rsfMRI	ROI (PFC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	44.4 to 80.9**
Hao et al. (2015)[3]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC, VC, SSC and CC combined)	Yes	DBaN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	48.9 to 72.7**
Plis et al. (2014)	HC = 191 SZ and FEP = 198	sMRI	WB voxel-level	No	DBN	HC vs. SZ	91**
Kim et al. (2016)[4]	HC =50 SZ =50	rsfMRI	WB region-level	Yes	SAE	HC vs. SZ	85.8
Munsell et al. (2015)	HC = 48 TLE = 70	DTI	WB region-level	No	SAE	HC vs. TLE	69.0
Yang et at. (2014)	HC =31 SCA2 = 4 SCA6 = 27 AT = 18	sMRI	ROI (Cerebellum)	No	SAE	HC vs. SCA2 vs. SCA6 vs. AT	86.3

[1] ADNI dataset; [2] CADDementia dataset, [3]ADHD-200 dataset; [4]COBRE dataset; *Sample sizes for the fine-tuning stage only (pre-training included an additional 386 samples); **F-score; ***Range of accuracies obtain from the different datasets used; HC, healthy controls; SZ, schizophrenia, FEP, first episode psychosis; ADHD, attention deficit/hyperactive disorder; ADHD-C, attention-deficit/hyperactive disorder combine subtype; ADHD-I, attention-deficit/hyperactive disorder inattentive subtype; ADHD-H, attention-deficit/hyperactive disorder hyperactive subtype; SCA2, spinocerebellar ataxia type 2; SCA6, spinocerebellar ataxia type 6; AT, ataxia-telangiectasia; TLE, temporal lobe epilepsy; AD, Alzheimer's disease; MCI, mild cognitive impairment; CC, cingulate cortex; VC, visual cortex, PFC, pre-frontal cortex; SSC, somatosensory cortex; sMRI, structural MRI; rsfMRI, resting-state functional MRI; CT, computed tomography; PET, Positron emission tomography; DTI, diffusion tensor imaging; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; AE, autoencoder, SAE, stacked autoencoder; FCC, fully-connected cascade; DBN, deep belief network, DBaN, deep Bayesian network; CNN,

convolutional neural network; DAE, deep autoencoder; DBM, deep Boltzman machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron; nr, not reported.

Attention-deficit/hyperactive disorder. With regards to attention-deficit/hyperactivity disorder (ADHD), all five studies included here have used resting-state neuroimaging data. For example, Deshpande et al. (2015) applied a fully connected cascade artificial neural network - a variation of the multilayer perceptron – to functional connectivity from ADHD and healthy controls. The model successfully distinguished between the inattentive and combined subtypes from healthy controls with an accuracy of 90% for both comparisons, while the two subtypes were discriminated with an accuracy of 95%. Connections between frontal areas and the cerebellum were identified as the most discriminating features. There is also evidence that healthy children and children diagnosed with three different ADHD subtypes (inattentive, hyperactive and combined) can be distinguished in one single model using a multiclass approach, without the need to perform binary classifications between healthy controls and each ADHD subtypes. This evidence comes from three studies that have used data from different sites taken from the ADHD-200 consortium, a data-sharing platform aimed at understanding the neural basis of ADHD (Milham, Fair, Mennes, & Mostofsky, 2012). Kuang et al. (2014) attempted to discriminate between healthy controls and ADHD subtypes (inattentive, hyperactive and combined) using data acquired from three different sites. Rather than looking at the whole brain, the authors first parcellated the brain and trained different DBNs for each brain area using the voxel-level intensities within a given region as features, to examine which part of the brain best discriminated ADHD (regardless of subtypes) from healthy controls. A 4-way DBN was then performed for the each best discriminating area – prefrontal (PFC), cingulate (CC) and visual (VC) cortex – in each one of the three datasets separately (dataset 1: PFC=37.4%, CC=37.1%, VC=34.4%; dataset 2: PFC=54.0%, CC=54.0%, VC=51.2%; dataset 3: PFC=71.8%, CC=72.7%, VC=68.8%). Kuang and He (2014) partially replicated these findings by applying the same deep learning approach to functional measures of the prefrontal cortex; this allowed a 4-way classification accuracy of 44.4%, 55.6% and 80.9% in three independent samples from the ADHD-200 consortium. Finally, Hao et al. (2015) used the same type of features to first identify the most discriminating areas – prefrontal, cingulate, somatosensory and visual cortex – and then combined them within a single model. The resulting

input data were put through a deep Bayesian network (DBaN), where a DBN was used to reduce the dimensionality of the data and a Bayesian network was used to extract the relationships between the data. The resulting model achieved a 4-way classification accuracy of 48.8%, 54.0% and 72.7% for three independent samples also taken from the ADHD-200 consortium. These three studies suggest that deep learning can be used to solve multiclass classifications problems, as all performances were well above chance level (25% for a classification with 4 classes). In addition, these studies suggest that deep learning can extract meaningful information from patterns of brain functioning to classify ADHD from controls and, more notably, to differentiate between ADHD subtypes. Nevertheless, we note that all four studies conducted in ADHD had unbalanced sample sizes between classes. For example, in Kuang et al. (2014), there were just between 2 and 5 children in the Inattentive subtype within each site, while the number of healthy children ranged from 69 to 110 per site. Similarly, each site in Kuang and He (2014) did not include any participants on at least one ADHD subtype which may have introduced a bias in the 4-way classification performed across all sites. With the exception of Hao et al. (2015) which reported sensitivity and specificity, all studies assessed model performance by estimating the overall accuracy. This metric is simply the proportion of participants correctly identified, and therefore does not take the unbalance between classes into account; this means that it is possible to have a good overall accuracy even if several participants from a class are misclassified (or even if all participants from a class are misclassified if the sample size for that class is very small compared to the total sample size). Therefore, given the highly imbalanced sample sizes, the possibility that the performances reported in these studies are inflated cannot be ruled out. This possibility is supported by the observation of much lower sensitivities (43.9%, 22.9% and 55.6% for each site) than specificities (68.8%, 87.7% and 83.0%), in Hao et al. (2015). In addition, by running several models (i.e. one for each region), authors may have increased the risk of finding a positive result by chance. Perhaps a more appropriate strategy for investigating multiple regions could have been using the whole set of regions, include a nested procedure to select the most informative ones within the training set, and only then test the model; alternatively, they could have also corrected their final results for the multiple regions/models.

Psychosis. With respect to psychosis, two studies have been performed with promising results.

Using structural MRI data from four independent studies, Plis et al. (2014) applied a DBN to the original preprocessed images obtaining an impressive F-score of 91%. While this was a highly promising result, the patients group included both first episode and chronic schizophrenia patients, which could have diluted the models' performance. More recently, Kim et al. (2016) extracted functional connectivity patterns obtained from resting-state functional MRI of individuals diagnosed with schizophrenia and healthy controls and performed a series of experiments with an SAE-based model, in which different hyperparameters were tested. The proposed model consisted of an SAE with weight sparsity control, i.e. only a random selection of neurons in a given layer was activated, that classified schizophrenia patients and controls with an accuracy of 85.5%, outperforming SVM by a margin of 8.1%. Consistent with the literature on brain functional abnormalities in schizophrenia (Minzenberg, Laird, Thelen, Carter, & Glahn, 2009), the most relevant features for the classification were the functional connectivity between the thalamus and the cerebellum, the frontal and temporal areas and between the precuneus/posterior cingulate cortex and the striatum. Despite this encouraging result, the sample sizes for each class were modest (50 for each group) and, therefore, it is not clear how well these findings will generalise to a different sample. Nevertheless, both studies suggest that deep learning can effectively classify psychosis patients on the basis of neuroanatomical and neurofunctional information. Despite the evidence that structural and functional data provide complementary information on the neural basis of psychosis (Cabral et al., 2016; Radua et al., 2012; Schultz et al., 2012), to date there have been no deep learning studies using a multimodal approach in psychosis. In addition, despite the evidence that psychosis, similar to AD, is preceded by a prodromal stage (Yung et al., 2005), there have been no studies applying deep learning to neuroimaging data to classify individuals at high risk of developing psychosis from healthy controls or distinguishing between high risk individuals who will and will not develop the illness.

Temporal lobe epilepsy. One study examined the potential of deep learning to classify healthy individuals and patients diagnosed with temporal lobe epilepsy (TLE) from diffusion-weighted images (DWI) (Munsell et al., 2015). A stacked autoencoder was used to extract meaningful features from patients' connectome while SVM was chosen as the classifier. Deep learning was suggested as an attractive machine learning alternative because it is capable of encoding latent,

nonlinear relationships in high dimension data. This combination yielded a relatively modest accuracy of 69%. In addition, this model was outperformed by another approach where features were extracted using a well-known linear automated method (ElasticNet) instead, which achieved an accuracy of 80%. This discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. Given the high complexity resulting from the numerous parameters to be estimated, deep learning models are more prone to overfitting (high performance on the training data while performing poorly on unseen data) than conventional machine learning approaches. One standard solution, that the authors did not use, is to address this issue by tuning the level of model complexity and penalizing highly intricate ones in order to have better generalizing models.

Cerebellar ataxia. One study was conducted in cerebellar ataxia (CA), a neurodegenerative disorder that affects mainly the cerebellum, with multiple genetics variations each with its characteristic pattern of anatomical degeneration. Yang et al. (2014) applied a stacked AE to T1-weighted images of the cerebellum taken from healthy controls and individuals suffering from three CA subtypes: spinocerebellar ataxia type 2 (SCA2), spinocerebellar ataxia type 6 (SCA6) or ataxia-telangiectasia (AT). The proposed method classified the four groups with an accuracy of 86.3%, an impressive result for a 4-way classification. However, the confusion matrix reported by the authors indicates that no case with the SCA2 subtype was correctly classified. Because the sample size of this group (only four participants) contributed very little for the total sample size (80), it is still possible to misclassify all its cases and achieve a low error rate. In such cases, a high accuracy can be misleading, as it may reflect an overestimation of the algorithm's performance (Arbabshirani et al., 2017). Balanced accuracy, for example, is a potentially useful alternative as it calculates the average of correct predictions of each class individually (Alberg, Park, Hager, Brock, & Diener-West, 2004).

In short, since the first study published in 2013, there is already preliminary evidence that deep learning allows the accurate classification of a range of neurologic and psychiatric disorders, by extracting discriminating features from either single or multimodal imaging as well as other types of data such as clinical and cognitive information.

4.3.2. Conversion to illness

From Mild Cognitive Impairment to Alzheimer Dementia. A total of 8 studies have attempted to predict transition to illness using neuroimaging data, and all of them have focussed on the transition from MCI to AD (Table 4.2). Most studies used a multimodality approach, with three of them also including clinical measures in the prognostic model. The highest accuracy (83.3%), was achieved by a model which included sMRI, PET, CSF and two clinical measures: the MMSE and the ADAS-cog (H.-I. Suk et al., 2015). Interestingly, the lowest performance (57.4%) resulted from a model which used the same input data (sMRI, PET, CSF, MMSE and ADASCog) and a similar sample size (Li et al., 2014). However, the two studies differed on the deep learning approach, with the former employing a semi-supervised approach with a multilayer perceptron pretrained using a stacked sparse autoencoder, and the latter using a pure supervised approach. These findings highlight the potential impact of the deep learning architecture on performance, although we cannot exclude the contribution of other sample-specific factors to the results (e.g. recruitment criteria). Overall, this initial sample of studies suggests that individuals diagnosed with MCI who later convert to dementia can be identified using cutting-edge deep learning methods. Although, in general, accuracies are not as high as when classifying AD or MCI from healthy controls, this is not surprising since brain differences as well as clinical and cognitive symptoms between those identified as being at risk who do and do not develop a disorder are likely to be subtle. In addition to these encouraging results, the suitability of deep learning to multiclass classification means this analytical approach can easily be employed to examine the biomarkers of different stages of the illness. Four studies have taken advantage of this by conducting 4-way classifications to discriminate between no eminent risk of AD (healthy controls), individuals in the prodromal stage who did not (MCI-C) and did develop dementia (MCI-C) and established Alzheimer's (AD). Accuracies ranged from 46.3% to 53.8%. By using a deep Boltzmann machine to extract features from structural MRI and PET images, Liu et al. (2015) classified the four groups with an overall accuracy of 53.8%. Suk et al. (2016) examined the replicability of a deep learning approach known as deep weighted subclass-based sparse multi-task learning (DW-S2 MTL) in two different datasets, considering both binary and multi-way comparisons. The proposed model, specifically designed to mitigate the effect of less useful features for classification, showed a

comparable performance for both binary (74.2% vs. 73.9%) and 4-way (53.7% vs. 47.8%) classifications, thus suggesting good replicability. Taken collectively, these studies provide initial evidence that deep learning methods could be used to discriminate amongst different stages of illness – a common challenge in standard clinical settings.

Table 4.2. Conversion to illness.

Authors, year	Sample size	Technique	WB voxel-level/ WB region-level/ ROI	Feature selection	DL architecture	Comparison	Acc (%)
Liu et al. (2015a)[1]	HC = 204 AD = 180 MCI-C=160 MCI-NC=214	sMRI	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	46.3
Suk et al. (2014)[1]	MCI-C = 76 MCI-NC =128	sMRI & PET	WB region-level	Yes	DBM	MCI-NC vs MCI-C	71.6
Liu et al. (2015a)[1]	HC = 77 AD = 85 MCI-C=67 MCI-NC=102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	53.8
Liu et al. (2014)[1]	HC =77 AD = 65 MCI-C= 67 MCI-NC = 102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	47.4
Suk et al. (2015b)[1]	MCI-C = 43 MCI-NC =56 AD =51 HC =52	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C AD vs MCI-C vs MCI-NC vs HC	74.2 53.7
	MCI-C = 167 MCI-NC =236 HC= 52 AD = 198	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C AD vs MCI-C vs MCI-NC vs HC	73.9 47.8
Li et al. (2014)[1]	MCI-C = 43 MCI-NC =56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	MCI-NC vs MCI-C	57.4
Suk and Shen (2013)[1]	MCI-C = 43 MCI-NC =56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	MCI-NC vs MCI-C	75.8
Suk et al. (2015a)[1]	MCI-C=43 MCI-NC=56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	MCI-NC vs MCI-C	83.3

[1] ADNI dataset; DL: deep learning; HC, healthy controls; AD, Alzheimer's disease; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; sMRI, structural MRI; PET, Positron Emission Tomography; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; SAE, stacked autoencoder; DBM, deep Boltzmann machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron.

4.3.3. Treatment outcome

Prediction of response to treatment is a research area of high clinical interest. In several psychiatric and neurological disorders, a better understanding of why some patients benefit from a certain treatment whereas others do not, could help clinicians make more-effective treatment decisions and improve long-term clinical outcomes (Mechelli, Prata, Kefford, & Kapur, 2015). However, so far, only one study has used deep learning to predict clinical response to treatment (Table 4.3).

Table 4.3. Treatment outcome.

Authors, year	Sample size	Technique	WB voxel-level/ WB region-level/ ROI	Feature selection	DL architecture	Comparison	Acc (%)
Munsell et al. (2016)	TLEns = 41 TLEs = 29	DTI	WB region-level	No	SAE	TLEns vs TLEs	57.0

DL: deep learning; HC, healthy controls; TLE-ns, temporal lobe epilepsy without seizures; TLE-ns, temporal lobe epilepsy with seizures; DTI, diffusion tensor imaging.

Munsell et al. (2015) attempted to develop an algorithm that distinguished between patients with TLE who did and did not benefit from surgical treatment. This was implemented using a stacked autoencoder to extract meaningful features from the connectome of patients who were then classified using SVM. This model, however, yielded a low accuracy of 57%. For comparison, the author investigated another option where features were extracted with an alternative linear approach instead of an autoencoder. This second model resulted in a higher accuracy of 70%. Again, this discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. This model comprised 4 layers, resulting in a high number of weights to be estimated which, together with a modest sample size (41 patients without seizures and 29 with seizures after treatment), might have resulted in overfitting.

4.3.4. How does deep learning compare to a traditional machine learning approach?

A total of twenty-five studies included in this review compared a deep learning model against a kernel-based model (SVM or MKL) in order to elucidate how deep learning compares to a more conventional machine learning approach. The results of these comparisons are shown in Figure

4.5. It can be seen that, for the majority of studies, deep learning showed improved performance compared to SVM. Given the small sample of studies, it is difficult to identify specific characteristics of the studies associated with greater or smaller improvement in performance following the implementation of deep learning.

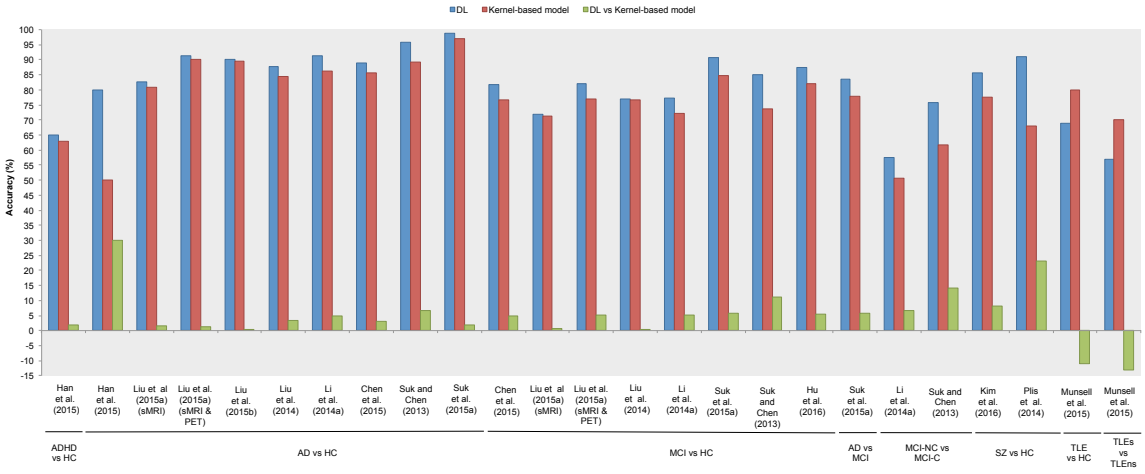


Figure 4.5. Results of studies comparing deep learning and kernel-based models. The graph shows the accuracies (F-score for Plis et al. (2014)) for deep learning models (blue), kernel-based models (red) and the difference between the two (green). HC, healthy controls; ADHD, attention deficit and hyperactive disorder; AD, Alzheimer's disease; MCI, mild cognitive impairment; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; SZ, schizophrenia; TLE, temporal lobe epilepsy; TLEs, temporal lobe epilepsy with seizures after treatment; TLEns, temporal lobe epilepsy without seizures after treatment.

However, a margin favouring deep learning studies appears to be more evident in studies that have integrated different modalities with cognitive and/or clinical data (Figure 4.6). This anecdotal observation is consistent with the notion that deep learning is a powerful tool for detecting abstract relations within the data, especially between different types of data that are likely to be associated in complex ways, such as neuroimaging and clinical/cognitive information (Plis et al., 2014). Since deep learning requires a large number of observations to learn increasingly complex patterns compared to conventional machine learning methods, one would expect to find a greater difference between the two methods as sample size increases. However, the effect of sample size on the difference in performance is unclear, possibly due to the small number of studies

currently available. There is a minority of studies where SVM/MKL matched or even outperformed the proposed deep learning model. Amongst these, Munsell et al. (2015) reported the largest margin favouring SVM. However, this article had one of the smallest sample sizes (118 for the diagnostic comparison and 70 for the treatment outcome comparison) while employing one of the deepest networks with 5 layers.

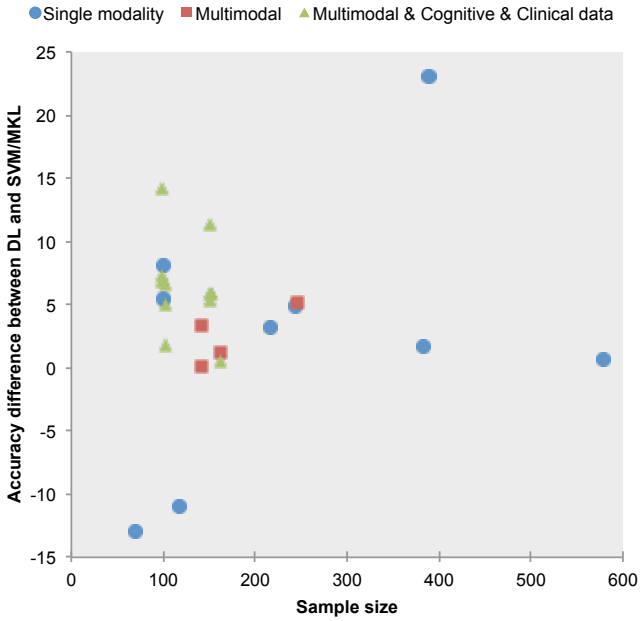


Figure 4.6. Difference in performance of deep learning against kernel-based methods for single modality, multimodal as well as for multimodal with cognitive/clinical data studies, according to sample size.

Notably, out of all the studies comparing the two approaches, Munsell et al. (2015) was the only one that did not make any formal attempt to prevent overfitting of the deep learning model, for example through the use of regularization. We note that susceptibility to overfitting becomes more pronounced when deeper and thus more complex networks are used, as in the study by Munsell et al. (2015), due to the higher number of weights to be estimated (Srivastava et al., 2014). Therefore, we speculate that the use of small sample sizes, coupled with the high-dimensionality of the data (i.e. when the number of variables highly exceeds the number of participants), may have increased the risk of overfitting in this study.

4.4. Discussion

Machine learning has been gaining considerable attention in the neuroimaging community due to

its advantages over traditional analytical methods based on mass-univariate statistics. In particular, machine learning methods take the inter-correlation between regions into account, while mass-univariate methods operate under the assumption that different regions act independently. In addition, machine learning methods can be used to make inferences at the single-subject level – a critical difference with mass-univariate analytical methods that are only sensitive to differences at group-level. Deep learning is a type of machine learning which is increasingly used in neuroimaging after leading to major scientific advances in the areas of speech recognition, computer vision and natural language processing by significantly outperforming other state-of-the-art classification methods (Krizhevsky et al., 2012; Le, 2013). There are two main characteristics that distinguish deep learning from conventional machine learning methods: first, deep learning is capable of learning features from the raw data without the requirement for *a priori* sophisticated feature extraction, resulting in a more objective or less bias-prone process; second, deep learning uses a hierarchy of nonlinear transformations, which make this approach ideally suited for detecting complex, scattered and subtle patterns in the data. Given its ability to detect abstract patterns from the data, deep learning can be considered a promising tool in neuroimaging, as most brain-based disorders are characterized by a scattered and diffused pattern of neuroanatomical and neurofunctional alterations (Plis et al., 2014). In previous sections of this review, we have described the most common deep learning architectures and have provided an overview of the studies that have applied deep learning to neuroimaging data to investigate psychiatric and neurological disorders. In this final section, we discuss the main themes that have emerged from the review of these studies. These will include (i) consistencies and inconsistencies in the existing literature (ii) the promise of CNNs, (iii) the issue of multiclass classification, (iv) how deep learning performs compared with conventional machine learning methods, (v) interpretability of deep learning in neuroimaging, (vi) the challenge of overfitting and (vii) technical expertise and computational requirements. We conclude by discussing possible directions for future research.

4.4.1. Main conclusions from the existing literature

The majority of published studies have been conducted in patients with MCI and/or AD; this may be explained by the availability of ADNI, a very large open-source dataset including thousands of

patients, to the neuroimaging community (Mueller et al., 2005a, 2005b). However, studies have also been conducted in other disorders including ADHD, psychosis, TLE and cerebellar ataxia. Taken collectively, the findings published so far suggest that deep learning can be applied to neuroimaging data, including both structural and functional modalities, to classify diagnostic groups from healthy individuals. Indeed, the performance of the classifiers has been consistently high, with several studies reporting accuracies above 95% for binary classifications between patients and controls (Deshpande et al., 2015; Hosseini-Asl et al., 2016; Payan & Montana, 2015; Sarraf & Tofighi, 2016; H.-I. Suk et al., 2015; H.-I. Suk, Lee, et al., 2016; H. Il Suk & Shen, 2013). Nevertheless, the application of a supervised model for diagnostic classification is arguably circular: since diagnostic labels in the training and testing datasets are predetermined through clinical examination, logic dictates that a perfect performance from a machine learning algorithm will simply mimic clinical assessment. Being able to predict a future diagnosis, or anticipate who will and will not benefit from a certain treatment, are questions of greater translational value in clinical practice. A total of 8 studies have applied deep learning to neuroimaging data acquired from individuals with MCI to predict subsequent transition to AD with promising results. For example, Suk et al. (2015) successfully predicted conversion from MCI to AD with 83.3% accuracy, after combining structural MRI and PET data. However, no studies have yet examined transition to illness in other psychiatric disorders with a prodromal phase, such as psychosis, even though we know that it is possible to distinguish between converters and non-converters using conventional machine learning (Pettersson-Yeo et al., 2013; Valli et al., 2016; Zarogianni et al., 2013). Overall, the best results were achieved in neurologic disorders. This is a prevalent finding across neuroimaging machine learning studies (Orrù et al, 2012) and likely due to the fact that psychiatric disorders do not have, as of yet, reliable diagnostic biomarkers (e.g. Prata, Mechelli, & Kapur, 2014), and therefore diagnosis rests purely on clinical interviews, leading to diagnostic labels with poor reliability (Regier et al., 2013) and biological validity (Insel et al., 2010; Jablensky, 2016; Kapur, Phillips, & Insel, 2012). To our knowledge only one study has used deep learning to predict treatment outcome. Munsell et al. (2015) achieved an accuracy of 57% when classifying TLE patients who did and did not suffer from seizures after surgical intervention. As discussed earlier, however, this modest result could potentially be explained by the absence of formal strategies to avoid overfitting of the deep learning model.

Deep learning is a very flexible approach, meaning that is it possible to combine different architectures and manipulate a range of hyperparameters within the same model. In addition, the vast majority of existing studies have been published in the last 2 years, and therefore the field of deep learning applied to neuroimaging of brain-disorders should be considered still at a very early stage. Possibly as a result of this combination of flexibility and novelty, the methodology of the studies reviewed in this article varied considerably. For example, some studies employed a whole-brain approach whereas others focussed on a subset of regions of interest; some studies used the raw data without any form of feature selection or dimensionality reduction whereas others performed a number of transformations on the data to extract relevant features; and different studies used different deep learning architectures. Such methodological variability means that, at present, the reliability and replicability of the existing results remain unclear.

4.4.2. The promise of convolutional neural networks

CNNs are a particular type of feedforward neural network inspired by how the human visual cortex process information. Over the past decade, CNNs have been breaking records in computer vision across several competitions, making this approach a very promising one (Krizhevsky et al., 2012). Consistent with this, our review has shown that CNNs have generated the most encouraging results in the context of neuroimaging. In its raw form, neuroimaging data comprises millions of voxels. Considering the current computational resources available, putting all voxel intensities through a fully connected network would lead to an unfeasible number of weights to be estimated. Two intrinsic properties of CNNs - weight sharing and local connectivity - result in a significantly reduced number of weights, making it computationally possible to run the network at the voxel-level. Although in neuroimaging CNNs have only been used to examine MCI and AD patients, the accuracies of the studies published so far have been consistently high (i.e. ≥95% for AD and ≥86% for MCI versus controls). High accuracies have been observed with different modalities including structural MRI (A. Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan & Montana, 2015), resting-state fMRI (Sarraf & Tofighi, 2016) and CT imaging (X. W. Gao & Hui, 2016), as well as with small (X. W. Gao & Hui, 2016; Sarraf & Tofighi, 2016) and large (A. Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan & Montana, 2015) sample sizes. Hosseini-Asl et al. (2016) used

an alternative and interesting approach which involved pre-training a CNN in one Alzheimer's dataset (CADDementia) and then fine-tuning and testing it in another dataset from the same diagnostic group (ADNI). The results were very promising for both 2-way and 3-way classifications (HC vs. AD; HC vs. MCI; AD vs. MCI; and HC vs. AD vs. MCI), although it should be noted that the ADNI sample was of modest size. Taken together, these results are in line with the successful performances of CNN-based models reported in other scientific areas, and highlight CNNs as a promising tool in neuroimaging.

4.4.3. From binary to multiclass classifications

In the context of neuroimaging, the vast majority of conventional machine learning studies have relied on binary classifications involving the comparison between a group of patients and a group of healthy controls (Orrù et al., 2012; Wolfers et al., 2015). This can be explained by the fact that these studies have typically employed SVM, which was originally designed for binary classification problems (Hsu & Lin, 2002). However, the real challenge for clinicians is not to differentiate between patients and controls but to develop biomarkers which could be used to choose amongst alternative diagnoses or different stages of illness progression. Looking forward, therefore, machine learning models will need to be able to discriminate amongst several possible alternatives in order to inform real-world clinical decision making. Many approaches have been proposed to enable SVM to handle multiclass classification problems (Fei & Liu, 2006; Hsu & Lin, 2002). However, this is still an active research area (Kumar & Gopal, 2011) and none of the proposed approaches have been tested in the context of neuroimaging. Most neuroimaging studies using SVM addressed the multiclass problem by performing several binary classifications (for example, AD vs. HC, MCI vs. HC and AD vs. MCI) or one-against-all classifications (for example, AD vs. MCI & HC and MCI vs. AD & HC). However, similar to other approaches such k-nearest neighbours and gaussian process for example, deep learning requires less technical effort to perform multiclass comparisons, and therefore could contribute for a solution to this issue. This is mainly due to the use of the so-called softmax function in the output layer, which can be considered an extension of the binary logistic regression to several classes. Here the output reflects the probability of belonging to each class, which is a more intuitive index of class membership than some of the most sophisticated indices being developed for SVM multiclass

solutions (Fei & Liu, 2006). In light of its suitability for multiclass classification, a number of studies have used deep learning to carry out 3 or 4-way classifications between different disorder subtypes or different stages of illness. For example, three of these studies were able to classify children into healthy controls and three ADHD subtypes (inattentive, hyperactive and combined) (Hao et al., 2015; Kuang, Guo, An, Zhao, & He, 2014; Kuang & He, 2014). Notably, there is also preliminary evidence for the use of deep learning to distinguish between individuals at no imminent risk of dementia, those identified at risk who will and will not develop dementia, and those with established Alzheimer's disease (Liu, Liu, Cai, Che, et al., 2015; Liu et al., 2014; H.-I. Suk, Lee, et al., 2016). These are encouraging findings, as they highlight how deep learning could help bridge the existing gap between neuroimaging findings and real-world clinical practice.

4.4.4. Is deep learning superior to conventional machine learning?

Despite the success of deep learning in several scientific areas, the superiority of this analytical approach in neuroimaging is yet to be demonstrated. On the one hand, deep learning has been described as a potentially more powerful approach than conventional shallow machine learning, as it is capable of learning highly intricate and abstract patterns from the data, which can particularly useful in the case of brain-based disorders (Plis et al., 2014). On the other hand, given that neuroimaging data is very high-dimensional, the nonlinear approach of deep learning might not be advantageous as there are not enough data points to extract meaningful nonlinear patterns from the data, whereas the linear approach employed in conventional shallow machine learning might be more appropriate. Here we tried to clarify this issue by systematically examining the difference in performance between deep learning and conventional shallow machine learning in studies which used both approaches. A total of twenty-five studies reported classification accuracy for both deep learning and conventional shallow machine learning, with the latter being a kernel-based method, either SVM or MKL. For the majority of these studies deep learning performed better than conventional shallow machine learning as shown in Figure 4.5, and in some cases the difference was by a reasonable margin (Xiaobing Han et al., 2015; Plis et al., 2014; H. Il Suk & Shen, 2013).

From the available evidence, it is not clear whether deep learning tends to perform better under

specific circumstances, for example depending on the modality type or the sample size. However, our systematic review provides anecdotal evidence that studies combining imaging and non-imaging data tend to have a larger margin in favour of deep learning (Figure 4.6). This is consistent with the notion that the association between brain abnormalities and cognitive symptoms, for example, is likely to exist at a deep and abstract level, and as such can be captured more effectively by deep learning methods than traditional shallow machine learning methods (Plis et al., 2014).

Overall, machine learning techniques thrive with larger samples. One would expect this to be especially true for deep learning: since a deep model is inherently more complex than conventional shallow machine learning models, larger sample sizes should be needed to compensate for the greater number of parameters to be estimated and to take full advantage of deep learning's ability to detect highly intricate and abstract patterns in the data. We were therefore expecting to see an increase in the margin by which deep learning outperforms kernel-based methods as sample sizes increase. Such increase however was not observed, as the pattern of difference in performance did not seem to vary systematically with sample size; one possibility is that larger sample sizes than those used in the existing literature would be required to detect increases in the margin by which deep learning outperforms kernel-based methods.

In conclusion, our review suggests that, overall, deep learning performs better than conventional shallow machine learning. In light of the increasing interest in deep learning, however, we cannot exclude a publication bias which favoured studies showing the superiority of this new analytical approach relative to conventional shallow machine learning methods (Boulesteix, Lauer, & Eugster, 2013). As the number of studies applying deep learning to neuroimaging data increases, a thorough assessment of publication bias would be useful to establish the reliability of this initial trend in favour of deep learning.

4.4.5. Interpretability of deep learning in neuroimaging

Despite having demonstrated state-of-the-art performances across several fields, deep learning has been under scrutiny for its lack of transparency during the learning and testing processes

(Alain & Bengio, 2016; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). For example, deep neural networks have been referred to as a “black box” in contrast with other techniques, such as logistic regression, which are less complex and more intuitive. Such lack of transparency has important implications for the interpretability of the results when deep learning is applied to neuroimaging data. Due to the multiple nonlinearities, it can be challenging to trace the consecutive layers of weights back to the original brain image in order to identify which features (e.g. regions) are providing the greatest contribution to classification (H.-I. Suk et al., 2015). This information however would be useful in the context of clinical neuroimaging where the aim is not only to detect but also localise abnormalities. A first potential issue is that a model with an excellent performance may be using irrelevant features (e.g. orientation of the images, imaging artefacts), as oppose to clinically meaningful information (e.g. regional grey matter, connectivity between different brain regions), to classify participants. A second potential issue is that an accurate model which provides no information about the underlying neuroanatomical or neurofunctional alterations would be of limited clinical utility, for example with respect to treatment development and optimization.

Despite its complex inner workings which make the visualization and interpretation of the weights challenging, deep learning *can* be used in a way which enables transparency. This is illustrated by several neuroimaging studies included in this review that did report the most important features (Deshpande et al., 2015; J. Kim et al., 2016; Liu et al., 2014; H.-I. Suk, Wee, et al., 2016). However, these studies used a variety of approaches to isolate the most informative features, and at present there is no standard and intuitive method for visualizing weights or interpreting latent feature representations (H.-I. Suk et al., 2015). This has motivated several attempts to develop new and intuitive ways of enhancing the interpretability of deep learning within the recent literature (Dahne et al., 2015; Grün, Rupprecht, Navab, & Tombari, 2016; Simonyan, Vedaldi, & Zisserman, 2013; Yosinski et al., 2015; Zeiler & Fergus, 2014). There are two main methodological approaches to address this issue, including input modification methods and deconvolution methods. Input modification methods are visualization techniques that involve the systematic modification of the input and the measurement of any resulting changes in the output as well as in the activation of the artificial neurons in the intermediate layers of the network. An example of

these methods is the so-called occlusion method (Zeiler & Fergus, 2014) which involves covering portions of the input image up to find the areas of the input data that influence the probability of the output classes. In contrast, deconvolution methods aim to determine the contribution of one or more features of the input data to the output. This involves selecting an activation of interest in an output neuron and then computing the contribution of each neuron in the next lower layers to this activation. Here a number of strategies are available to model the nonlinearities present across the layers, for example, deconvnet (Zeiler & Fergus, 2014) and guided backpropagation (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014).

4.4.6. The challenge of overfitting

Overfitting is arguably one of the main challenges in machine learning. Given their inherent complexity, deep learning networks are particularly prone to overfitting, i.e., learning irrelevant fluctuations in the data that limit generalizability. Not surprisingly, different approaches to address this issue, known as regularization strategies, have been developed and are now present in most deep learning algorithms. In section 4.2.1.4 we described some of the most commonly used regularization strategies applied to modern deep learning, namely weight decays and dropout. As expected, several studies reviewed here have used some form of regularization. The majority (Hosseini-Asl et al., 2016; J. Kim et al., 2016; Liu, Liu, Cai, Che, et al., 2015) have employed the L1 or L2 norms, which prevent overfitting by penalizing very low or very high weight values. At least one study (Li et al., 2014) employed dropout, where a random number of nodes and respective connections are temporarily removed to extract different sets of features that can independently produce a useful output. The importance of regularization strategies in deep learning could potentially account for the fact that Munsell and colleagues, who trained 4- and 5-hidden layer models (for inferring diagnostic and treatment outcome, respectively) without using any form of regularization, reported such low performance for deep learning (Munsell et al., 2015).

An additional approach for minimising the risk of overfitting involves reducing the dimensionality of the data before inputting them into the model. A possible way of achieving this is by extracting region- or patch-level features (as opposed to using voxel-level data). Using different types of features (whether voxel, patch or region) can have implications for how detailed the information

inputted into the model is (for example, voxel-level features are very detailed, and also very noisy; region-level features on the other hand, ignore more localized patterns and are less sensitivity to noise). Another option to reduce dimensionality is feature selection. Feature selection is common in conventional machine learning, where linear methods such as principal component analysis, independent component analysis or elastic net, are used to select the most discriminating features that are then fed to a classifier. However, the use of conventional feature selection methods prior to a deep learning model seems counterintuitive, since one of the main advantages of deep learning is the ability to learn, through a purely data-driven method, the most useful features for classification. Several studies reported in this review have attempted to reduce the dimensionality of the data by extracting region- or patch-level features, using feature selection, or combining the two approaches. We note, however, that all CNN-based models were applied to voxel-level data without being preceded by any form of feature selection and yet reported consistently high performances on unseen data. This suggests that deep learning, and CNN-models and particular, can perform well with neuroimaging data without the requirement to downsize or even preprocess the data. For example, Hosseini-Asl et al. (2016) achieved high levels of accuracy when classifying AD and healthy controls, after applying a CNN to voxel-level data without any preprocessing or even skull stripping of the images. This finding has potential implications for the development of clinical tools, as it suggests that it might be possible to apply deep learning to raw neuroimaging data, thereby saving time as well as technical resources.

In addition to overfitting due to the complex nature of deep learning, overfitting may have also occurred in several of the studies presented here due to the inappropriate use of feature selection and dimensionally reduction (e.g. PCA) approaches. From most articles, it was not clear whether these strategies were implemented within a cross-validation framework. Failure to do so means that the train and test data are no longer independent, and therefore, when the model is tested the data is not completely new to the model as it should be, resulting in inflated performances. This issue has also been highlighted in other review articles in psychiatric neuroimaging (Arbabshirani, Plis, Sui, & Calhoun, 2017; Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015).

4.4.7. Technical expertise and computational requirements

The studies reviewed in this article employed a wide range of deep learning architectures and hyperparameters. Such flexibility is what makes deep learning a very powerful tool but comes at a potentially high cost. The number of layers, the number of nodes within each layer and the activation function of each node are only a few examples of a long list of variables one has to consider when designing and optimizing a deep learning model. Automated optimization strategies are not yet widely available, making optimisation a manual process that requires a great deal of technical expertise and is potentially prone to subjective bias. Since the number of parameters to be estimated is very large, the computational requirements of deep learning are also more demanding than those of conventional machine learning methods. For example, Kim et al. (2016) reported that the estimation of a deep learning model with three hidden layers took 100 times longer than the estimation of a standard SVM model (~3.3 days vs. 0.8h). However, with the fast-growing availability of graphical processing units (GPUs), the application of deep learning to neuroimaging data is likely to become less and less time-consuming in the future.

4.4.8. Limitations of deep learning

Despite the promising initial findings, there are some important limitations that need to be considered. Perhaps the most obvious (and perhaps most controversial) limitation is the need for very large sample sizes to train deep learning models. This is due to their highly level of complexity, i.e. a large number of observations is needed in order to make up for the substantial number of parameters to be estimated. However, large sample sizes have been one the biggest challenges in psychiatric neuroimaging research, where a sample with a few hundred participants is considered exceptionally large. This is in sharp contrast with disciplines where deep learning is considered state-of-the-art and where data is abundant. For example, areas in which deep learning has excelled, such as image recognition, typically use datasets with one million examples (Krizhevsky, Sutskever, & Hinton, 2012). Significant larger samples, with dozens of thousands of participants, are now emerging in neuroimaging research (Kaufmann et al., 2019) which may allow explore deep learning closer to its full potential. Another important and related limitation of deep learning models is the higher risk of overfitting compared to traditional machine learning models. Indeed, despite the use of regularization, deep learning models are, due to its complex

nature, more prone to learning noise or fluctuations in the training data that are not necessarily related to the task and therefore hinder generalizability. The actual training of deep learning models is also challenging. The optimization of the weights is typically formulated as a highly non-convex optimization problem, with multiple local minima that make it very difficult to find a global minima (Goodfellow, Bengio, & Courville, 2016), although many suggest this may not be a serious problem (LeCun, Bengio, & Hinton, 2015). Training also requires extensive expertise to understand how the vast number of hyperparameters effect the learning process (e.g. learning rate, number of neurons, number of layers) and how best to tune them. Finally, also rather controversial is the lack of interpretability of deep learning models. At the moment, the strength of deep learning lies mostly on its pure data-driven mechanistic prediction and therefore traditional models may be more appropriate for gaining insight into neurobiological mechanisms of psychiatric and neurological disorders.

4.5. Conclusions and Future Directions

While still in its initial stages, the application of deep learning in neuroimaging has shown promising results and has the potential of leading to fundamental advances in the search for imaging-based biomarkers of psychiatric and neurologic disorders. Nevertheless, several improvements will be required before the full potential of deep learning in neuroimaging can be achieved. Firstly, given the complexity of deep learning models, we need to move away from studies with small to modest sample sizes in favour of much larger cohorts. A possible way of achieving this is through multi-centre collaborations, in which data is collected using the same recruitment criteria and scanning protocols across sites. A further way of increasing the sample size is through multi-site data sharing initiatives, such as ADNI for Alzheimer's disease and ADHD-200 for ADHD. Secondly, the integration of CNN and recurrent neural networks (i.e. networks that allow the processing of data with sequential inputs such as videos or speech) is likely to lead to significant advances in deep learning in the next few years (Donahue et al., 2017). In neuroimaging, this integration could be particularly useful for analysing fMRI data, as it would allow the detection of intricate spatial patterns while simultaneously modelling the temporal component of the BOLD signal. Thirdly, we anticipate that an increasing number of neuroimaging studies will make use of transfer learning, which involves using previously learned features from

a large sample of similar enough images. This could help tackle the curse of dimensionality – a common problem in neuroimaging studies of brain disorders (A. Gupta et al., 2013; Hosseini-Asl et al., 2016). Evidence from vision science, where deeper models such as VGG net (Simonyan & Zisserman, 2014), residuals networks (He, Zhang, Ren, & Sun, 2016) and Inception-v4 (Szegedy, Ioffe, Vanhoucke, & Alemi, 2016) are achieving the highest performances, suggests that transfer learning could be particularly useful when deeper models are employed. Fourthly, we suggest that the so-called augmentation technique - which it is commonly used in computer vision – could be useful in the context of neuroimaging. This technique involves increasing the sample size by applying transformations to the data (e.g., rotation, shear, scaling), and then train a model that is invariant to such transformations. The use of augmentation could also address the issue of modest sample sizes and lead to a decrease in prepossessing time (because steps such as rotation may become redundant). Finally, the use of deep learning to predict continuous scores is another interesting area for further research with potential clinical applicability, following the encouraging results obtained using conventional machine learning methods (Gong et al., 2014; Stonnington et al., 2010; Tognin et al., 2013). So far, only one study has used deep learning to predict clinical scores from structural MRI scans in patients with Alzheimer's disease (Brosch & Tam, 2013).

In conclusion, the capacity of deep learning models to learn complex and abstract representations through nonlinear transformations, makes this a promising approach to single subject prediction in neuroimaging. While there are still important challenges to overcome, the findings reviewed here provide preliminary evidence supporting the potential role of deep learning in the future development of diagnostic and prognostic biomarkers of psychiatric and neurologic disorders.

Chapter 4 supplementary materials

sTable 4.1. Method for comparison and features used for studies comparing DL with another method.

Study	Comparison method	Features[1]
Han et al. (2015)	PCA + SVM	Voxel-level intensities extracted from resting-state fmri images
Liu et al. (2015a) (sMRI only)	Elastic net + SVM (LIBSVM)	83 ROIs of grey matter volume extracted from T1-weighted.
Liu et al. (2015a) (sMRI + PET)	Elastic net + MKSVM (rbf kernel) (LIBSVM)	83 ROIs with grey matter volume and regional average of cerebral metabolic rate of glucose extracted from T1- weighted and PET images, respectively.
Liu et al. (2015b)	Elastic net + MKSVM (rbf kernel) (LIBSVM)	83 ROIs with grey matter volume and regional average of cerebral metabolic rate of glucose extracted from T1- weighted and PET images, respectively.
Liu et al. (2014)	Elastic net + MKSVM (rbf kernel) (LIBSVM)	83 ROIs with grey matter volume and regional average of cerebral metabolic rate of glucose extracted from T1- weighted and PET images, respectively.
Li et al. (2014a)	PCA + LASSO + SVM (linear kernel)	189 total features containing the grey matter volume of 93 ROIs extracted from T1-weighted images, regional average of cerebral metabolic rate of glucose extracted from PET images, and three CSF biomarkers (Aβ42,t−tau, and p-tau)
Suk & Chen (2013)	MKSVM (linear kernel)	93 ROIs with grey matter volume and regional average of cerebral metabolic rate of glucose extracted from T1- weighted and PET images, respectively, and three CSF biomarkers (Aβ42,t−tau, and p-tau)
Suk et al. (2015a)	MKSVM (linear kernel)	93 ROIs with grey matter volume and regional average of cerebral metabolic rate of glucose extracted from T1- weighted and PET images, respectively, and three CSF biomarkers (Aβ42,t−tau, and p-tau)
Hu et al. (2016)	SVM	90×90 functional connectivity matrix
Kim et al. (2016)	SVM (linear kernel) (LIBSVM)	Whole-brain functional connectivity matrix containing 6670 pairs of regions from the AAL atlas.
Plis et al. (2014)	SVM (rbf kernel)	60465 voxel-level grey matter volumes extracted from T1-weighted images
Munsell et al. (2015)	Elastic net + SVM (linear kernel) (LIBSVM)	82×82 density connectivity matrix

1Used for both the DL and comparison method unless specified.

Chapter 5

Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence

This chapter is based on the paper entitled Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence published in Schizophrenia Bulletin.

Vieira, S., Gong, Q., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-Garcia, V., Setién-Suero, E., Scheepers, F., van Haren, N. E. M., Kahn, R. S., Reis Marques, T., Murray, R., David, A., Dazzan, P., McGuire, P. & Mechelli, A. (2019). Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophrenia bulletin*.

5.1. Introduction

Over the last three decades, traditional mass-univariate neuroimaging approaches have revealed neuroanatomical abnormalities in individuals with psychosis (Chan et al., 2011; Fusar-Poli et al., 2011; Smieskova et al., 2010; Torres et al., 2016; A Vita et al., 2012). Because these abnormalities were detected using group-level inferences, it has not been possible to use this information to make diagnostic and treatment decisions about individual patients. Machine learning is an area of artificial intelligence that promises to overcome this issue by learning meaningful patterns from the imaging data and using this information to make predictions about unseen individuals (Davatzikos et al., 2005). Several machine learning studies have attempted to use neuroanatomical data to distinguish patients with established schizophrenia from healthy individuals, with promising results (Kambeitz et al., 2015; Orrù et al., 2012; Wolfers et al., 2015; Zarogianni et al., 2013). At present, however, there are two important limitations in the existing literature that limit the translational applicability of the findings in real-world clinical practice. First, given the well-established effects of illness chronicity and anti-psychotic medication on brain structure (Bora et al., 2011; Navari & Dazzan, 2009; van Erp et al., 2016, 2018; Antonio Vita et al., 2015), it is unclear to what extent classification was based on neuroanatomical changes associated with these factors rather than the onset of the illness per se. Consistent with this, both disease-stage and anti-psychotic medication were identified as significant moderators in a recent meta-analysis of diagnostic biomarkers in schizophrenia (Kambeitz et al., 2015). Also in line with this, Pinaya et al (2016) reported that the same machine learning model that was able to distinguish between patients with established schizophrenia and healthy controls with an accuracy of 74%, showed poor generalizability (56%) when applied to a cohort of individuals with first episode psychosis (FEP). Taken collectively, these findings suggest that representations learned from patients with established schizophrenia may not be applicable to individuals with a first episode of the illness. Second, the clinical utility of any machine learning-based diagnostic tool for detecting patients with an established illness is likely to be very limited; in contrast, detecting the initial stages of an illness, when diagnosis may be uncertain and treatment is yet to be decided, is likely to have much greater clinical utility.

So far only a limited number of studies have applied machine learning to neuroanatomical data

in the initial stages of the illness when the effects of illness chronicity and anti-psychotic medication are minimal. These studies have produced inconsistent results, including poor (e.g. 51% in Winterburn et al (2017)), modest (e.g. 63% in Pettersson-Yeo et al (2013)) and good (e.g. 86% in Borgwardt et al (2013) or 85% in Xiao et al (2017)) accuracies. There are a number of possible reasons for such inconsistency. First, most of the studies used small samples (N≤50) (see Kambeitz et al (2015) for a meta-analysis), which have been shown to yield unstable results (Nieuwenhuis et al., 2012; Schnack & Kahn, 2016). Second, the vast majority of studies used data from a single site, and as such may have generated results that were specific to the characteristic of the local sample rather than the illness per se. Third, a series of recent articles have highlighted potential methodological issues that may have caused inflated results in some of the published studies (Arbabshirani et al., 2017; Janssen, Mourão-Miranda, & Schnack, 2018; Schnack & Kahn, 2016; Winterburn et al., 2017; Wolfers et al., 2015; Woo et al., 2017). These issues include, for example, (i) failure to use a nested cross-validation (CV) framework to avoid *knowledge-leakage* between training and test sets; (ii) failure to perform feature transformation and/or selection within a rigorous CV framework resulting in so-called “double dipping”; (iii) publication bias leading to an over-representation of positive findings, especially in studies with small samples and (iv) failure to test performance on additional independent samples. Also, we note that all studies have employed traditional ‘shallow’ machine learning techniques, such as support vector machine and logistic regression. The intuitiveness of such techniques has made them very popular in neuroimaging studies of psychiatric and neurological disease. Deep learning is an alternative type of machine learning which has been gaining considerable attention in clinical neuroimaging (Arbabshirani et al., 2017; Pinaya et al., 2016; Vieira, Pinaya, & Mechelli, 2017; Wolfers et al., 2015). Contrary to traditional machine learning, where the immediate input data is used to extract patterns (hence the term ‘shallow’), deep learning learns complex latent features of brain structure through consecutive nonlinear transformations (hence the term ‘deep’) which are then used for classification. Given its ability to learn more intricate and abstract patterns, deep learning might be particularly suitable to detect the subtle and heterogenous neuroanatomical abnormalities characteristic of the early stages of psychosis (Chan et al., 2011; Plis et al., 2014; Schnack, 2017).

This study aims to elucidate the extent to which the application of machine learning to neuroanatomical data allows distinction between patients with first episode psychosis and healthy controls at the individual level. To overcome the limitations of previous studies, we used a total of five datasets from different sites, each with a sample size above the recommended threshold for a stable performance (Nieuwenhuis et al., 2012), and employed both shallow and deep machine learning techniques. In addition, following a series of recent articles highlighting potential methodological issues in the existing literature (Arbabshirani et al., 2017; Janssen et al., 2018; Schnack & Kahn, 2016; Winterburn et al., 2017; Wolfers et al., 2015; Woo et al., 2017), we put in place a series of precautions to minimise the risk of overfitting. Based on previous studies, we hypothesize that: 1) FEP and HC will be classified with statistically significant performances ranging between 70% and 80% (Kambeitz et al., 2015) and 2) deep learning will perform better than traditional shallow approaches (Vieira et al., 2017).

5.2. Methods

5.2.1. Participants

Participants were recruited as part as five independent studies carried out in multiple sites, all of which have been previously published:

- Site 1: Chengdu, China (Gong et al., 2015)
- Site 2: London, England (GAP study) (Di Forti et al., 2009)
- Sites 3 and 4: Santander A and B, Spain (PAFIP study) (Pelayo-Terán et al., 2008)
- Site 5: Utrecht, The Netherlands (GROUP study) (Korver et al., 2012)

All patients were experiencing their first psychotic episode, defined as the first manifestation of psychotic symptoms meeting criteria for a psychotic disorder, as specified by the DSM-IV (APA, 2000) or ICD-10 (Organization World Health, 1992). Recruitment details are reported in Chapter 2, sections 2.1.

5.2.2. MRI data acquisition and preprocessing

At all 5 sites, volumetric MRIs were acquired using a T1-weighted protocol. At four sites, the scanner field strength was 3T, and at 2 sites it was 1.5T. The details of the image acquisition

sequence are reported in Chapter 2, section 2.2.2. From each image, three types of data features were extracted (see Chapter 5 supplementary material):

- Voxel-wise grey matter volume (VWGMV): whole-brain voxel-wise estimate of the local density of GM in a given voxel region (Ashburner, 2007).
- Voxel-wise cortical thickness (VWCT): cortical thickness maps in which each voxel in the grey matter is assigned a thickness value (Hutton et al., 2008, 2009).
- Surfaced-based regional volumes and cortical thickness (SB-ROIs): volume and thickness of predefined cortical and subcortical regions extracted with FreeSurfer (Fischl, 2012).

5.2.3. Statistical analysis

5.2.3.1. Demographic and clinical variables

Differences in age, sex and total intracranial volume (TIV) between FEP and healthy controls (HC) were examined using an independent-samples t-test and chi-square test, as implemented in the Statistical Package for the Social Sciences 24.0 (SPSS 24.0).

5.2.3.2. Group-level comparisons

For completeness, a standard group-level analysis was also carried out for each site and type of feature set separately. See Chapter 5 supplementary material sections 1.4.1. and 2.1 for methods and results, respectively.

5.2.3.3. Multivariate pattern recognition analysis

5.2.3.3.1. Dimensionality reduction: principal component analysis

Principal component analysis (PCA) was used to reduce the number of voxels of the VWGMV and VWCT maps (see Chapter 5 supplementary material).

5.2.3.3.2. Classifiers

Four methods were used for classification: k-nearest neighbours (KNN), logistic regression (LR), support vector machine (SVM) and deep neural networks (DNN). These methods were chosen based on their increasing order of complexity: KNN is a straightforward algorithm, whilst deep

learning can be more powerful at the expense of transparency; popularity: SVM and LR and among the most machine learning techniques used in previous studies; and novelty: deep learning has yielded promising results in psychiatric neuroimaging but is yet to be applied to FEP. KNN, LR and SVM were implemented using the Scikit-Learn library (Pedregosa et al., 2011) (sklearn) for python 3.5. Deep neural network (DNN) was implemented using Tensorflow v.1.4 (Abadi, Chu, et al., 2016) and Keras v.2.1 (Chollet & others, 2015) libraries. The random seed was kept the same for all models to ensure the reproducibility of the results. This approach guaranteed that the starting weights and train/test split at each fold of the CV would remain the same within and between algorithms for the same site.

5.2.3.3.2.1 K-nearest neighbours

K-nearest neighbours (KNN) is a non-parametric method based on multivariate pairwise distance measures between data points. Once presented with unseen data, it calculates the Euclidean distance between this new data point and each of the surrounding neighbours. Classification is done by assigning the unseen data to the same class as the majority of its neighbours (Altman, 1992). The optimal number of neighbours was tuned via grid search by testing 10 possible odd values ranging from 3 to 21 in increments of 2.

5.2.3.3.2.2. Logistic regression

Logistic regression (LR) was implemented via elastic net, a regularized regression that combines the regularizations L1 and L2 penalties of LASSO (Least Absolute Shrinkage and Selection Operator) and ridge regression, respectively. While the ridge penalty retains all variables and minimizes the impact of irrelevant features, the LASSO penalty discards unimportant variables (H. Zou & Hastie, 2005). Grid search was used to find the optimal relative contribution of each penalty via tuning of the hyperparameter l1_ratio as defined by sklearn from eleven possible values between 0 and 1 with increments of 0.1.

5.2.3.3.2.3. Support vector machine

Support vector machine (SVM) is a supervised machine learning technique that maps the input data into a feature space using a set of similarity functions known as kernels. In this feature space,

the model finds the optimal separating hyperplane by finding the largest margin of separation between the two classes within the training set. Once the hyperplane is determined, it can be used to predict the class of new unseen observations (Pereira & Mitchell, 2008; Vapnik, 1995). In this study, a linear kernel was chosen to contrast with the characteristic nonlinear approach of deep learning. The soft margin (C) parameter, that controls the trade-off between having zero training errors and allowing misclassifications, was tuned from a possible range of values (2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}) using grid search, i.e. all possible values in a given range were tested.

5.2.3.3.2.4. Deep neural network

Given its flexible architecture, deep learning can be used to build a variety of different neural networks (LeCun et al., 2015). Here we employed a deep neural network, with the components resulting from the PCA (for the VWGMV and VWCT data) or the SB-ROIs as inputs; the general-purpose deep neural network (DNN) was chosen as it allowed for automated and non-biased optimization of the hyperparameters, which in turn helps prevent overfitting. Deep neural networks are multi-layered fully-connected networks where higher-level features are learned as a nonlinear combination of lower-level features, thus allowing the extraction of complex and abstract patterns from the data. Once the model learns these higher-level features, it can determine a separation surface to classify the different classes (LeCun et al., 2015; Vieira et al., 2017). The performance of DNN models relies on the specification of several architectural and learning hyperparameters. To prevent bias, the number of layers, number of units, optimizer, learning rate, decay, activation function and epoch and were optimized using random search as implemented by sklearn. To decrease the chances of overfitting, two additional parameters were also included at each layer: i) L2 regularizer, which penalizes high weights (Krogh & Hertz, 1992) and ii) dropout, where randomly selected neurons are ignored during training (Srivastava et al., 2014). Each layer was initialized via Glorot (also known as Xavier) initialization (normal distribution) (Glorot & Bengio, 2010). In the output layer, the classification was performed by a softmax function. Training was carried out using a mini-batch with 8 training samples for VWGMV and VWCT data, and 128 for the SB-ROIs. DNN models were optimized via random search due to the high number of parameters to test: at each fold, 500 different combinations of randomly selected values for each parameter were tested. Table 5.1 shows all the possible values for each parameter.

Table 5.1. Parameters for tuning the DNN.

Parameter	Values
Number of layers	2, 3, 4, 5
Number of units	10, 20, 50, 75, 100, 150
Activation function	ReLU, Leaky ReLU
Learning rate	0.001, 0.005, 0.01, 0.1, 0.2
Learning rate decay	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}
Epochs	50, 100, 150
Optimizer	Stochastic gradient descent (SGD), Adam
Momentum	0.99, 0.9, 0.95
L2 coefficient	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}
Drop-out rate	0.2, 0.5, 0.7

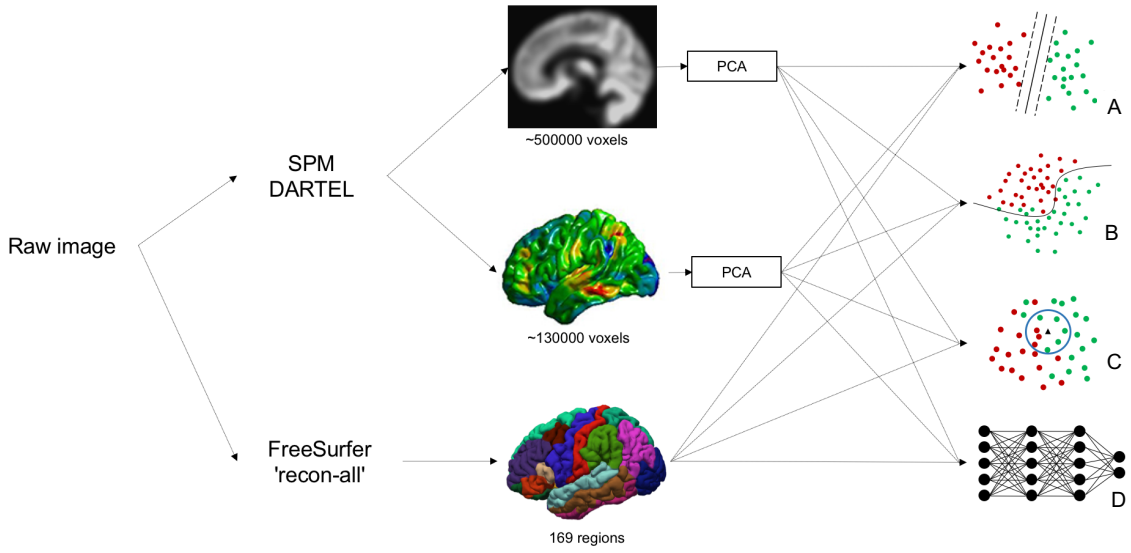


Figure 5.1. Analysis pipeline. Three features were extracted from each image: VWGMV, VWCT and FreeSurfer SB-ROIs. The dimensionality of VWGMV and VWCT was reduced through PCA. The resulting features were analysed with four classifiers: **A.** SVM, **B.** LR, **C.** KNN and **D.** DNN.

5.2.3.3.3 Model training and testing

Within-site classification. All models were assessed through a nested 10-fold stratified CV framework (Figure 5.2) to ensure that the data for hyperparameter tuning and the data to test the algorithm were strictly independent. A 10-fold CV was chosen as a trade-off between bias, variance and the demanding computational resources required to run the DNNs.

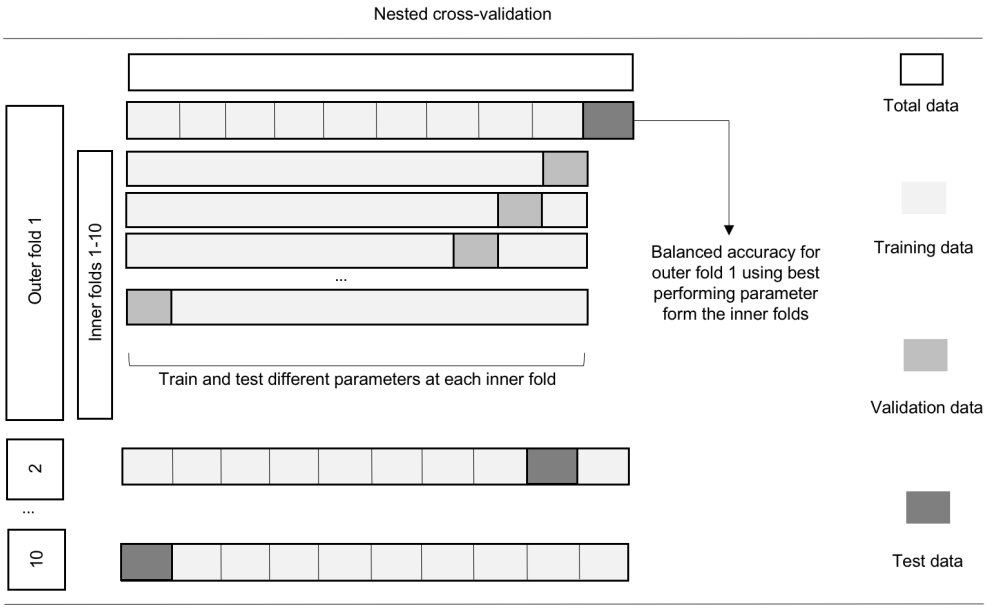


Figure 5.2. Schematic representation of nested CV. Nested CV involves a secondary inner CV loop using the training data from the primary outer CV split, where different sets of hyperparameters are tested (e.g. different values for the C parameter for SVM). The best performing hyperparameters amongst the 10 inner folds are then used to train a model in the whole training set defined by the outer loop. This model is then tested using the test set of the outer loop. The final performance is estimated by averaging accuracies in the test set across all 10 outer folds.

Cross-site classification. The best site-level model was further tested in each one of the remaining independent samples. This was done by running the 10 instances of the trained model (one trained model for each one of the 10 CV folds) on the independent sample. This resulted in ten sets of class membership probabilities, one for each of the 10 instances. The final predicted label was estimated using the soft voting method, i.e. by averaging the ten predicted probabilities (participants with a probability equal or higher than 0.5 were assigned to the patients' group, otherwise they were assigned to the healthy controls group).

5.2.3.3.4. Performance measures

Balanced accuracy, sensitivity and specificity as defined in section 2.3.2.5.1. in Chapter 2 were used to measure the performance of each classifier. Statistical significance of the balanced accuracy was determined by permutation testing with 1000 permutations (see Chapter 5 supplementary material).

5.2.3.3.5. Effect of anti-psychotic medication and psychotic symptoms

To examine whether anti-psychotic medication and psychotic symptoms contributed to the classifiers' performance, chlorpromazine equivalents and positive and negative psychotic symptoms were regressed against the predicted labels using a logistic regression as implemented by the Logit function from the statsmodel python library. Because all patients from site 1 were anti-psychotic naïve, the investigation of the effects of medication was limited to sites 2, 3, 4 and 5. The size of the effects of medication and psychotic symptoms was measured in terms of odds ratio (OR) and respective 95% confidence interval (CI). The statistical significance threshold was set to 0.05.

5.3. Results

5.3.1. Socio-demographic and clinical parameters

No statistically significant differences were identified between patients and controls for age, sex or total GM volume at each site (Table 5.2).

5.3.2. Single-subject classification

Can we detect FEP at the individual level?

Balanced accuracy, sensitivity, specificity and statistical significance for each feature set of interest and site are presented in Table 5.3 (for a visual display of the accuracies and standard deviations see sFigure 5.3 in the supplementary materials). Overall, results were poor to modest across all types of feature sets and sites; although the site with the smallest sample size (site 2) showed the lowest performance consistently across all feature sets. Overall, regression analyses examining the effect of anti-psychotic medication and psychotic symptoms on the performance of each classifier did not show a significant effect (see Chapter 5 supplementary materials).

What is the most effective type of feature set?

There was no clear effect of type of feature set across sites. However, it can be seen that SM-ROIs data tended to yield higher accuracies, especially when analysed with DNN.

Can we generalise the results from one site to the others?

The best performances were achieved by two DNN models at sites 1 and 3 using SB-ROIs, with 70.5% and 70.2%, respectively. However, both models generalized poorly when tested on the remaining sites: specifically, the DNN model from site 1 achieved accuracies (sensitivity/specificity) of 52.1% (56.3%/47.9%), 61.1% (70.0%/52.7%), 52.1% (65.7%/38.6%) and 50.0% (48.3%/51.7%) when applied to sites 2 through 5, respectively; whilst the DNN model from site 3 achieved accuracies of 52.2% (96.5%/8.4%), 49.2% (83.5%/33.4%), 55.1% (70.1%/40.0%) and 51.0% (67.5%/34.6%) when applied to sites 1, 2, 4 and 5, respectively. To examine the possibility that poor generalizability was due to site differences, the same DNN model was applied to the total data with the five sites added as additional features. Features weights were then investigated to determine the importance of site. Results showed that out of the 174 features, the weights for site 1, 2, 3, 4 and 5 ranked 110, 150, 108, 71 and 112, respectively.

5.4. Discussion

In the last few years, there has been increasing interest in the translational potential of machine learning approaches in psychosis. As the field matures, there is emerging scepticism about replicability and generalizability, which has led to recent calls for greater caution in the interpretation of the findings (Arbabshirani et al., 2017; Schnack & Kahn, 2016; Winterburn et al., 2017; Wolfers et al., 2015; Woo et al., 2017). This study aimed to elucidate the extent to which the application of machine learning to neuroanatomical data allows detection of individuals at the early stages of psychosis when the effects of illness chronicity and anti-psychotic medication are minimal. To overcome the limitations of the existing literature, we used five independent datasets and put in place a series of methodological precautions to avoid overoptimistic results. Contrary to expectation, the performances of all methodological approaches tested were poor to modest across all sites. Below we discuss some of the main aspects that emerge from our investigation, including sample size, full independence of training and test data, cross-site generalizability and testing multiple pipelines. We conclude the discussion by considering possible future directions.

Table 5.2. Demographic and clinical characteristics for FEP and HC for each site.

		Chengdu, China (N=224)		London, England (N=142)		Santander A, Spain (N=220)		Santander B, Spain (N=210)		Utrecht, The Netherlands (N=162)	
		HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP
N		112	112	71	71	110	110	70	140	81	81
Sex (%)	M	51 (46)	51 (46)	36 (51)	36 (51)	68 (62)	68 (62)	45 (64)	90 (64)	64 (79)	64 (79)
	F	61 (54)	61 (54)	35 (49)	35 (49)	42 (38)	42 (38)	25 (46)	50 (46)	17 (21)	17 (21)
		χ2=ns		χ2=ns		χ2=ns		χ2=ns		χ2=ns	
Age M(SD)		27.2 (7.3)	25.7 (8.1)	26.8 (7.1)	26.4 (6.2)	29.7 (7.8)	28.5 (8.6)	27.3 (7.5)	28.3 (7.6)	26.9 (8.0)	25.2 (5.9)
		t=ns		t=ns		t=ns		t=ns		t=ns	
TIV (L) M(SD)		1.5 (0.1)	1.5 (0.2)	1.5 (0.2)	1.5 (0.2)	1.5 (0.1)	1.4 (0.2)	1.5 (0.1)	1.5 (0.1)	1.6 (0.1)	1.5 (0.2)
		t=ns		*t*=ns		*t*=ns		*t*=ns		*t*=ns	
Positive symptoms M(SD)		-	24.6 (6.6)[a]	-	13.9 (5.5)[a]	-	14.7 (4.6)[b]	-	14.4 (4.1)[b]	-	15.9 (6.3)[a]
Negative symptoms M(SD)		-	18.2 (7.7)[a]	-	16.0 (6.0)[a]	-	6.3 (4.6)[c]	-	6.1 (5.0)[d]	-	16.2 (6.9)[a]
Duration of illness (years) Med (IQR)		-	0.3 (1.1)	-	1.1 (0.3)	-	0.3 (0.7)	-	0.3 (0.9)	-	0.6 (1.0)

TIV: total intra-cranial volume; L: liters; M: male; F: female; FEP: first episode psychosis; HC: healthy controls; PANSS: Positive and Negative Symptoms Scale; [b]SAPS: Scale for the Assessment of Negative Symptoms; [c]SANS: Scale for the Assessment of Negative Symptoms; ns: $p>.0$

Table 5.3. Accuracies (sensitivity/specificity) for each feature set and algorithm across all sites using nested 10-fold stratified cross-validation. The classifier yielding the best balanced accuracy is highlighted in bold for each site.

		SB-ROIs	VWGMV	VWCT
Site 1 Chengdu China	KNN	60.7** (74.3/47.1)	**60.7** (49.5/71.9)**	62.1** (72.1/52.1)
	LR	61.9** (64.9/58.9)	60.1** (62.9/58.6)	**67.2** (65.8/68.5)**
	SVM	61.3** (66.4/56.2)	**60.7** (63.0/58.5)**	52.7* (24.6/97.3)
	DNN	**70.5** (72.2/68.8)**	57.7** (59.5/56.0)	66.4** (63.9/68.3)
Site 2 London UK	KNN	56.7 (50.9/62.5)	43.9 (33.6/54.3)	53.5 (38.4/68.6)
	LR	51.6 (45.0/58.2)	51.9 (53.8/50.0)	**61.6** (63.2/60.0)**
	SVM	45.9 (49.3/42.5)	**53.9 (53.4/54.3)**	51.0 (96.3/5.7)
	DNN	**58.8* (49.5/68.0)**	40.8 (47.4/34.3)	53.4 (52.4/55.3)
Site 3 Santander A Spain	KNN	59.6** (45.5/73.6)	50.5 (31.8/69.1)	58.0* (50.0/66.4)
	LR	58.6* (58.2/59.1)	63.2** (63.6/62.7)	59.1* (58.2/60.0)
	SVM	60.5** (61.8/59.1)	**65.9** (68.2/63.6)**	51.8* (90.9/12.7)
	DNN	**70.2** (70.0/70.4)**	50.2 (52.7/63.6)	**59.6 (60.0/59.1)**
Site 4 Santander B Spain	KNN	56.6* (91.8/21.4)	58.9** (70.7/47.1)	59.5* (67.7/51.1)
	LR	54.8 (73.9/35.7)	59.6** (57.8/61.4)	**62.6** (56.8/62.4)**
	SVM	56.0 (65.0/47.1)	57.4* (71.9/42.9)	58.4* (71.9/52.9)
	DNN	**62.0** (76.8/47.1)**	**59.3* (81.4/37.1)**	58.8** (62.4/53.1)
Site 5 Utrecht The Netherlands	KNN	52.7 (53.6/51.8)	54.5 (33.8/75.3)	52.2 (36.5/67.9)
	LR	58.5* (61.7/55.4)	61.3** (56.8/65.7)	**60.5** (60.6/60.4)**
	SVM	**60.7** (59.7/61.7)**	**62.4** (63.1/61.8)**	56.3 (51.2/61.4)
	DNN	54.9 (59.2/51.8)	58.0** (58.1/57.9)	60.1** (56.1/64.2)

**p<.01, *p<.05; SVM: support vector machine; LR: logistic regression; KNN: k-nearest neighbours; DNN: deep neural network; SM-ROIs: surfaced-based regional volumes and cortical thickness; VWGMV: voxel-based morphometry; VWCT: voxel-based cortical thickness

5.4.1. Sample size, homogeneity and publication bias

A possible explanation for why our accuracies are lower than those reported in the existing literature is that some of the previous studies may have reported overoptimistic results due to the

use of fairly small sample sizes. To illustrate this possibility, we tested for an association between sample size and classification accuracy across studies using machine learning and structural MRI (sMRI) in the existing literature (see Chapter 5 supplementary material). Unsurprisingly, we found a moderate negative association for studies that examined established schizophrenia (r=-.41) and FEP (r=-.59; after excluding Xiao et al (2017) which was a clear outlier; Figure 5.3A). This is consistent with the notion that some of the previous studies may have reported overoptimistic accuracies due to the use of inadequate sample size.

There are at least two possible ways in which inadequate sample size can lead to an inflated estimation of the accuracy of an algorithm, including sample homogeneity and publication bias (Schnack & Kahn, 2016; Woo et al., 2017). Firstly, smaller samples tend to be more homogeneous, making it easier for an algorithm to learn shared abnormalities in patients relative to controls and resulting in higher accuracies. In contrast, larger samples tend to be more heterogeneous due to the loosening of inclusion criteria; in this case, it may be more challenging to find a shared pattern of abnormalities resulting in lower performances. This inverse relationship between sample size and accuracy was not observed in our investigation; however, this might be explained by the fact that there was not sufficient variability in sample size across our five datasets. Secondly, smaller samples tend to be unstable and thus yield underestimated as well as overestimated accuracies (Nieuwenhuis et al., 2012; Varoquaux, 2017). This may, in turn, lead to publication bias, with overestimated accuracies being more likely to be published. In their meta-analysis of machine learning studies of schizophrenia, Kambeitz et al (2015) reported that no publication bias was evident when all studies - including sMRI, fMRI and DTI - were examined together. To test for publication bias in sMRI studies, we repeated the same statistical analysis focusing on this modality (see supplementary material). This revealed a statistically significant asymmetry in the funnel plot of published studies, indicating the presence of publication bias (Figure 5.3B). This is in line with emerging concerns about possible over-representation of inflated performances in the literature (Arbabshirani et al., 2017; Schnack & Kahn, 2016; Winterburn et al., 2017; Woo et al., 2017).

5.4.2. Full independence of training and testing set data

Following recent recommendations on how to overcome methodological issues that may have led to initial inflated results (Arbabshirani et al., 2017; Wolfers et al., 2015; Woo et al., 2017), we adopted two important methodological precautions. First, the use of simple CV, in which the same test data is used to both tune model hyperparameters and evaluate its performance, has been criticized as it almost certainly leads to inflated performances (Arlot & Celisse, 2010; Varma & Simon, 2006). In the present investigation, algorithms were trained and tested via nested CV. This ensured that the test set remained fully independent from the training set, with only the latter being used to optimise model parameters. Second, implementing feature selection in a two-step approach, where for example univariate tests (e.g. t-test) are applied in the whole sample and only the statistically significant features are used for classification, is likely to result in overoptimistic performances as features are chosen based their performance on data that should be completely independent for testing the classifier. In the present investigation, therefore, transformations to the data such feature selection were implemented embedded within the CV framework, i.e. parameters were derived from the training data only and subsequently applied to the test set. The adoption of these methodological precautions, aimed at ensuring full independence between training and test data, might explain the fact that accuracies in the present investigation were lower than expected. Nevertheless, despite the precautions used to ensure the independence of the train and test, the voxel-level features used in this study were computed using a study-specific template (i.e. DARTEL) based on the total data available for each site. This is the most common approach in machine learning studies that use voxel-level psychiatric neuroimaging data (Kambeitz-Ilankovic et al., 2019; Pettersson-Yeo, 2013; Valli et al., 2016). However, it should be acknowledged that this is not the optimal approach for a machine learning study, since both the training and test data are transformed together using information present on both sets, and therefore compromising the independence between training and test sets. The decision to use this approach outside the CV framework was based on limited computational resources and time. A more rigorous solution would be to build a template using the training data only and apply it to both training and test data, at each iteration of the CV.

5.4.3. Cross-site generalizability

The use of independent samples to develop and validate an algorithm is a critical requirement if the ultimate aim is to develop flexible machine learning-based tools that could be used in a clinical setting (Arbabshirani et al., 2017; Woo et al., 2017). However, only a minority of studies have attempted to do this (Dluhoš et al., 2017; Schnack & Kahn, 2016; Schnack et al., 2014), and most of them have reported considerably lower performances in the independent sample. In the present investigation, the highest accuracies – obtained using specific combinations of dataset, type of feature set and algorithm – were 70% (in sites 1 and 3 with surface-based regional features and DNN); this performance would appear to be in line with previous similar studies. However, selectively reporting these accuracies from our wider set of results would have portrayed a distorted picture of the potential of machine learning to detect the initial stages of psychosis at the individual level (Janssen et al., 2018). This is especially true since, after testing these two models in independent datasets, their performance did not hold up indicating low cross-site generalizability. Such low cross-site generalizability could be due to site-related differences in scanning parameters, cultural interpretation of diagnostic criteria and ethnicity; therefore, it might be possible to achieve higher cross-site generalizability by combining samples that are homogenous with respect to these variables. Nevertheless, our current results indicate that algorithms developed using data from a specific centre do not perform well when applied to data from other centres, and thus have limited clinical applicability.

5.4.4. Testing multiple pipelines

Because existing studies tend to differ with respect to several methodological aspects, at present, it is difficult to say which pipeline is optimal for detecting FEP (Salvador et al., 2017). Multi-pipeline studies have therefore been proposed as a useful way to disentangle what aspects works best (Arbabshirani et al., 2017). Importantly, this approach may also help build more generalizable models, as the development of a bespoke, and possibility overfitted, pipeline to a local sample is less likely to occur. Consistent with this, Salvador et al (2017) tested the performance of a range of machine learning approaches in different types anatomical features extracted from patients with schizophrenia and controls, and reported lower accuracies (66-68%) compared to previous similar studies using a single pipeline. Winterbourn et al (2017) also used multiple pipelines in

FEP and reported poor to modest accuracies, ranging from 51% to 73%. Taken collectively, evidence from these studies, including our own, suggest that when features are not manually carved to fit one algorithm applied to one specific small dataset performance tends to drop. This can be seen in Figure 5.3A where two generations of studies emerge: initially there were mostly small single -site, -feature and -algorithm high-performance studies; more recently the use of 1) larger samples, 2) multi-centre studies (Dluhoš et al., 2017; Rozycki et al., 2018), 3) assessment of different algorithms and/or features in one/several site(s) (Salvador et al., 2017; Winterburn et al., 2017) or 4) independent sample testing (Dluhoš et al., 2017; Rozycki et al., 2018) are reshaping the original, and possibly over-inflated, enthusiasm with more realistic performances.

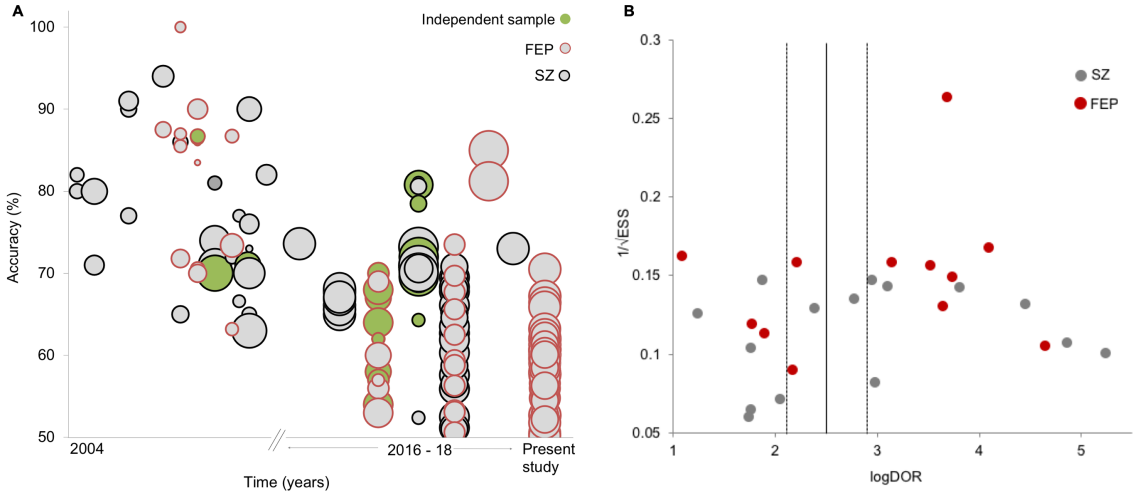


Figure 5.3. Summary of sMRI machine learning studies over time and funnel plot. **A**. Accuracy of diagnostic sMRI machine learning studies over time and sample size (circle increases with sample size). From the first study until 2015, the vast majority of studies reported accuracies ranging between 70% and 100%; from 2016, however, performances have dropped overall with accuracies ranging between chance-level and 85%. **B.** Funnel plot for sMRI studies in SZ and FEP showing the distribution of individual studies according to their sample size (1/√ESS) and effect size (logDOR). The plot revealed a statistically significant asymmetric distribution around the result of the meta-analyses of machine learning-sMRI studies (Kambeitz et al., 2015) (p=.013), indicating a bias favouring higher effect sizes.

5.4.5. What next for machine learning-sMRI studies of psychiatric disease?

Unlike group-level analysis, where larger samples lead to increased chance of detecting a

statistically significant result (even with a small effect size), in neuroimaging machine learning studies it has been observed that larger samples do not necessarily equate to better results; instead, these tend to lead to lower accuracies potentially due to increased heterogeneity (Schnack, 2017; Schnack & Kahn, 2016). In fact, this observed inversed relationship between sample size and performance is also not commonly observed in more classical machine learning applications (Banko & Brill, 2001). Despite this challenge, larger samples are likely to be more representative of the illness, less likely to overfit and thus carry more translational potential. Future machine learning studies will have to address this issue to overcome the increasingly apparent bottleneck in the performance that is arising with larger sample sizes (Figure 5.3A). A possible way of doing so could be to use normative models, where an individual is mapped against a normative model that should encompass the heterogeneity characteristic of the normal population. Here, illness is considered an extreme case within a normal range, which is likely to be a more ecologically valid approach than the traditional case-control paradigm (Marquand, Rezek, Buitelaar, & Beckmann, 2016; Sato, Rondina, & Mourão-Miranda, 2012).

Greater methodological standardization based on ‘good-practice recommendations’ could also help disentangle the current conflicting evidence. For example, guidelines for minimum sample size such as the threshold (n>130) proposed by Nieuwenhuis et al (2012) are a good start. The need for independent sample testing has also been widely acknowledged as an essential step towards generalizability (Arbabshirani et al., 2017; Woo et al., 2017) however, even the most recent studies do not always perform this. Moving forward, this type of generalizability test is likely to become a gold standard for machine learning diagnostic studies. More transparency in the implementation of machine learning is also needed. Several studies do not provide enough information about how the algorithm was trained and tested (Arbabshirani et al., 2017; Schnack, 2017; Tandon & Tandon, 2018). This hinders a thorough assessment of the validity of the study as well as its replicability. It should also be noted that, even if sMRI was able to distinguish between patients with FEP and disease-free individuals with high levels of accuracy, this would be of limited clinical utility. This is because, from a clinical translation perspective, the real challenge is not to distinguish between patients and disease-free individuals, but to develop biological tests that could be used to choose between alternative diagnoses and optimise

treatment (Tandon & Tandon, 2018).

5.5. Conclusion

The present investigation attempted to overcome the limitations of the existing literature using a number of strategies. Firstly, we studied FEP patients in which the effects of anti-psychotic medication and illness chronicity are likely to be minimal. Secondly, the sample size of each of our five datasets was greater than the recommended threshold for achieving a stable performance in machine learning-sMRI studies (Nieuwenhuis et al., 2012). Third, critical methodological precautions (e.g. nested CV and appropriate use of feature selection) were adopted to ensure an unbiased assessment of performance. Fourth, we systematically assessed the performance of a range of algorithms and features across several datasets, thereby minimizing the possibility of developing a bespoke and likely overfitted model to a single site. Fifth, we assessed the cross-site generalizability of the best models at the single-site level. Our findings suggest that the use of machine learning and sMRI allows detection of FEP at the individual level with relatively modest accuracies – lower than what was expected based on previous studies and much lower than what would be required for clinical translation. We speculate that some of the previous results may have been over-optimistic due to a combination of small sample sizes, less-than-rigorous methodologies and possible publication bias and argue that the current evidence for the diagnostic value of machine learning and structural neuroimaging should be reconsidered towards a more cautious interpretation.

The DNN models showed a small superiority compared to traditional approaches when ROIs were used as input features. This small improvement is in line with the growing popularity of deep learning in psychiatric neuroimaging (Durstewitz, Koppe, & Meyer-Lindenberg, 2019; Vieira, Pinaya, & Mechelli, 2017). However, when the significant additional computational costs and expertise required to achieve it are put in context, the superiority of DNNs is not that clear. For example, each DNNs used in this study took roughly 3 days to run. On the other hand, an SVM model with the same number of features took less than 10 minutes. The same has been observed by Kim et al (2016), where their DNN models applied to functional connectivity matrices took approximately 100 times longer to train compared to an SVM model. Nevertheless, with the

increasing availability of better computational resources, it likely that this discrepancy will decrease significantly in the next few years.

Over the past few years the number of machine learning studies in psychosis has been increasing rapidly (Tandon & Tandon, 2018). As larger samples and more powerful computational resources become available, this momentum is likely to continue to grow over the coming years (Bzdok & Yeo, 2017). Therefore, it is as important for the research community to be aware of the challenges and limitations of applying machine learning to psychosis including, for example, several potential ”distortion” of the findings along the machine learning pipeline as discussed in a recent review (Tandon & Tandon, 2018). In light of these challenges and limitations, the extent to which the application of machine learning in psychosis will lead to a more valid construct of the illness remains an open question. We encourage researchers to continue pursuing the integration of machine learning and neuroimaging, whilst exercising caution to avoid inflated results and ultimately a distorted view of the potential of this approach in psychiatric neuroimaging.

Chapter 5 supplementary materials

5.1.sMethods

5.1.1. Participants

5.1.1.1. Matching

sTable 5.1. Sample size of each dataset. We report the number of subjects available (top row), the number of subjects excluded after matching patients and controls for age and sex (middle row) and the final number of subjects included in the statistical analysis (bottom row).

	Site 1 Chengdu, China	Site 2 London, England	Site 3 Santander A, Spain	Site 4 Santander B, Spain	Site 5 Utrecht, The Netherlands
Available	330	204	257	223	225
Excluded	106	62	37	13	63
Final	224	142	220	210	162

5.1.2. MRI preprocessing

After checking all T1-weighted images for scanner artefacts and gross anatomical abnormalities, images were preprocessed to extract three types of anatomical features: voxel-wise GM volume (VWGMV), voxel-wise cortical thickness (VWCT) and surface-based volumes and cortical thickness (SB-ROIs).

5.1.2.1. Voxel-wise maps

Two different voxel-wise features were extracted: grey matter volume and cortical thickness. Common to both features, images were first reoriented along the anterior-posterior commissure line and set the anterior commissure as the origin of the spatial coordinates to assist the normalization algorithm. Reoriented images were then segmented into GM, WM and CSF partitions as implemented in SPM12 (Ashburner & Friston, 2005) (http://www.fil.ion.ucl.ac.uk/spm).

5.1.2.1.1. Grey matter volume

The segmentation tissue maps for each site were preprocessed separately using the DARTEL

toolbox (Ashburner, 2007). This procedure warps the grey matter and white matter partitions into a new study-specific reference space representing an average of all the subjects included in the analysis, thus maximizing accuracy and sensitivity (Scarpazza, Tognin, Frisciata, Sartori, & Mechelli, 2015; Yassa & Stark, 2009). The warped grey matter partitions were then affine-transformed into MNI space. An additional modulation step was used to scale the grey matter probability values by the Jacobian determinants of the deformations, thereby ensuring that the total amount of grey matter in each voxel was conserved after registration (Mechelli et al., 2005). Finally, the GM probability maps were smoothed using a standard 8mm FWHM Gaussian kernel.

5.1.2.1.2. Cortical thickness

A voxel-based Laplacian method (Jones et al., 2000), implemented as an SPM toolbox (Hutton et al., 2008, 2009), was used to create a voxel-based cortical thickness (VBCT) map for each subject using the GM, WM and CSF partitions generated in the segmentation step. Briefly, the resulting VBCT maps contained cortical thickness values within voxels identified as grey matter and zeros outside the cortex. Each VBCT map was warped into the corresponding site-specific DARTEL reference space. The warped images were then normalized to MNI space and smoothed with a 6 mm Gaussian kernel. The same warps, modulation and smoothing were also applied to a binary mask created from each original VBCT map. Subsequently the warped, scaled and smoothed VBCT maps were divided by the corresponding warped, scaled, and smoothed mask.

5.1.2.2. Surface-based volume and cortical thickness

FreeSurfer 5.3 (http://surfer.nmr.mgh.harvard.edu) (Fischl, 2012) was used to parcellate each participant's raw brain image into subcortical and cortical regions according to the Desikan-Killiany atlas (Desikan et al., 2006) using the 'recon-all' command. FreeSurfer is a well-established automated procedure for imaging preprocessing and analysis which details have been extensively described elsewhere (Dale, Fischl, & Sereno, 1999a; Fischl et al., 2002, 2004). A total of 169 features were used, including 33 volumes of subcortical structures plus volume and thickness of 34 cortical regions per hemisphere (after removing white matter hypo-intensities, 5th ventricle, optic chiasm and bilateral vessels and choroid plexus).

5.1.3. sStatistical analysis

5.1.3.1. Group-level analysis

5.1.3.1.1. Grey matter volume and cortical thickness

Voxel-based morphometry (VBM) was used to calculate group-level differences in VWGMV and VWCT between FEP and HC groups at each site. An independent-sample t-test was used with statistical inferences made at $p<0.05$ after family-wise error (FWE) correction for multiple comparisons and a minimum extent threshold of 5 voxels.

5.1.3.1.2 Surface-based regional volumes and cortical thickness

Mean difference between the FEP and HC groups for each SB-ROI was analysed with an independent-sample t-test as implemented in SPSS 24.0 using a statistical threshold of $p<0.05$ and additional Bonferroni correction for multiple comparisons.

All reported results (sResults, section 5.2) were obtained without covariates of no interest to ensure consistency between group- and individual-level statistical analyses. However, statistical analyses with age and sex as covariates were also carried out for completeness; this yielded identical results except for surface-based regional volumes and cortical thickness data (sTable 5.7-5.9).

5.1.3.2. Multivariate pattern recognition analysis

5.1.3.2.1. Dimensionality reduction: principal component analysis

Principal component analysis (PCA) is a well-established unsupervised method for feature reduction in neuroimaging. PCA reduces dimensionality by geometrically projecting the data into lower dimensions called principal components (PCs), with the aim of finding the best summary of the data using a limited number of PCs. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated features into a set of values of uncorrelated features (PC). PCs are then ranked according to explained variance in descending order. A detailed description of PCA is given elsewhere (Jolliffe, 2002; Lever et al., 2017). In the present investigation, PCA was mainly used to allow the processing of VWGMV and VWCT data with the general purpose DNN. As a fully connected network, the number of parameters to estimate would not be feasible

to compute without first reducing the dimensionality of the data. As regularized methods, LR and SVM would not require this step. However, PCA was also used in combination with these methods to facilitate comparison between all approaches as well as to alleviate computational requirements. PCA was implemented within the CV framework; at each fold, dimensionality was reduced by 1) extracting the minimum number of principal components whilst retaining cumulative 90% of the variance from the data in the training set only, 2) projecting all grey matter/cortical thickness maps onto the resulting principal components and 3) using the resulting values for classification and 4) projecting the test data into the same components derived from the training set, and using the former for testing. PCA was implemented using the default parameters of the class PCA from the decomposition module of the sklearn library (v0.20) (Pedregosa et al., 2011).

5.1.3.2.2. Feature scaling: Standardization

Standardization was performed by removing the mean and scaling to unit variance using StandardScaler from the preprocessing module of the sklearn library (v0.20) (Pedregosa et al., 2011). This procedure was applied to each feature independently. Standardization is a common requirement for many machine learning methods, since algorithms might behave poorly if the individual features do not resemble normally distributed data. In addition, features with bigger scales might dominate the loss function of the training algorithms. To avoid “double dipping”, the statistics (mean and variance) were obtained using only the training set, and these same values were used in the standardization of test set.

5.1.3.2.3. Most contributing brain regions of the DNN models

The most contributing were identified with the function SmoothGrad (Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017) as implemented by the iNNvestigate library (Alber et al., 2018). Briefly, SmoothGrad works by first adding Gaussian noise to several copies of the input data. Each copy is then put through the trained model and a saliency map is generated from the network’s gradients. This results in several saliency maps that are then average to estimate a final smoothed saliency map. Smoothgrad takes two parameters: noise level or standard deviation of the Gaussian perturbations, and n, the number of samples to average over. Here we use the default parameters, standard deviation of 0.1, and 64 copies of in the input.

5.1.3.2.4. Significance testing

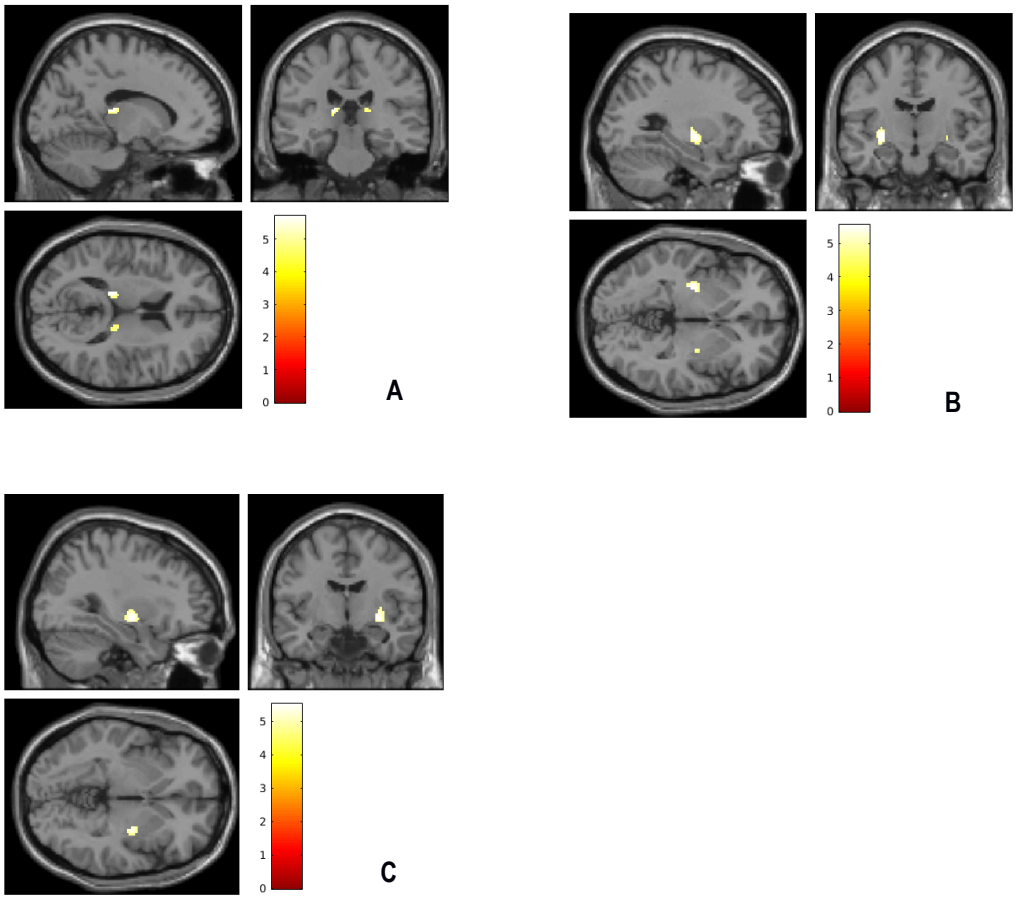
The balanced accuracy of each classifier was tested for significance using permutation testing, whereby subjects were randomly assigned to one of the classes (patients/control), so that the labels no longer match the data in any meaningful way, and the 10-fold CV cycle repeated 1000 times. This resulted in a distribution of accuracies reflecting the null hypothesis that the classifier did not exceed chance. The number of times the classifier's performance was greater than or equal to the true accuracy was divided by 1000 to determine a *p*-value. A *p*-value lower than 0.05 was considered statically significant.

5.2. sResults

No significant VWGMV decreases in FEP relative HC were found at any site. In contrast, VWGMV increases were detected in the bilateral thalamus at site 3; in the left putamen and the right pallidum at site 4; and in the right putamen at site 5. No significant increased or decreased VWCT was observed in FEP compared to HC at any site, except for site 1 in which FEP showed increased thickness in the left fusiform gyrus and left superior frontal gyrus. Significant differences in SB-ROIs between FEP and HC were found for sites 3 and 4. At site 3, patients showed smaller right hippocampus volume as well as a reduced thickness of the inferior parietal lobe; whereas at site 4 patients showed a significant cortical thinning in the left inferior temporal gyrus, pars opercularis and rostral middle frontal gyrus, as well as a larger 3rd ventricle. These results are presented in detail in sTables 5.2-5.7.

sTable 5.2. Group-level analysis: GWGMV.

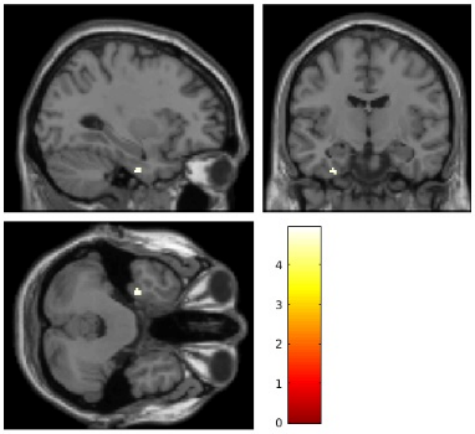
Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 3				
Left thalamus	-16,-28,12	37	5.5	.006
Right thalamus	16,-24,14	17	4.7	.014
Site 4				
Left putamen	-30,-12,-4	118	5.3	.001
Right pallidum	20,-10,-4	17	4.6	.014
Site 5				
Right putamen	30,-8,-4	85	5.3	.001



sFigure 5.1. Regions with increased GWGMV in FEP relative to controls in site 3 (A,) 4 (B) and 5 (C).

sTable 5.3. Group-level analysis: VWCT.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 1				
Left fusiform gyrus	-30,-10,-36	11	4.7	.012
Left superior frontal gyrus	-10,58,36	11	4.5	.012



sFigure 5.2. Cortical region with increased VWCT in FEP relative to controls in site 1.

sTable 5.4. Group-level analysis: SB-ROIs.

Region	t	*p*
FEP < HC		
Site 3		
Right hippocampus	4.3	<.001
Left inferior parietal (thickness)	3.9	<.001
Site 4		
Left inferior temporal gyrus (thickness)	3.6	<.001
Left pars opercularis (thickness)	4.0	<.001
Left rostral middle frontal gyrus (thickness)	4.8	<.001
FEP > HC		
Site 4		
Third ventricle	-4.1	<.001

sTable 5.5. Group-level analysis controlling for age and sex: VWGMV.

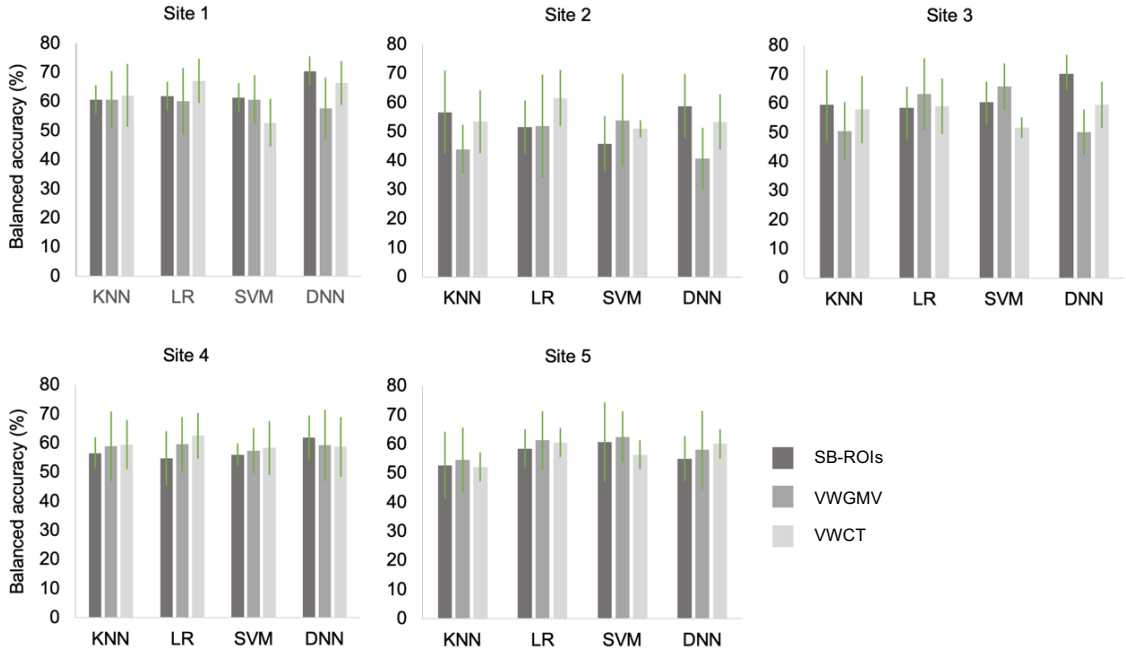
Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 3				
Left thalamus	-16,-28,12	40	5.6	.006
Right thalamus	16,-24,14	19	4.8	.014
Site 4				
Left putamen	-30,-14,-2	110	5.1	.001
Right pallidum	28,-10,-4	17	4.6	.014
Site 5				
Right putamen	30,-8,-6	70	5.2	.002

sTable 5.6. Group-level analysis controlling for age and sex: VWCT.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 1				
Left fusiform gyrus	-30,-10,-36	16	5.0	.008
Left superior frontal gyrus	-10,58,36	11	4.6	.012

sTable 5.7. Group-level analysis controlling for age and sex: SB-ROIs.

Region	F	*p*
FEP < HC		
Site 3		
Right hippocampus	22.4	<.001
Left inferior parietal gyrus (thickness)	20.7	<.001
Left precuneos (thickness)	15.3	<.001
Left superior frontal gyrus (thickness)	12.8	<.001
Left supramarginal gyrus (thickness)	17.2	<.001
Site 4		
Left parsopercularis (thickness)	15.6	<.001
Left rostral middle frontal gyrus (thickness)	21.9	<.001
Site 5		
Left hippocampus	15.7	<.001
FEP > HC		
Site 3		
Left lateral ventricle	14.8	<.001
Site 4		
Third ventricle	17.2	<.001



sFigure 5.3. Balanced accuracies and standard deviations of the different algorithms and feature sets for each site. KNN: k-nearest neighbours; LR: logistic regression; SVM: support vector machines; DNN: deep neural network; SB-ROIS: surfaced-based regional volumes and cortical thickness; VBM: voxel-based morphometry; VBCT: voxel-based cortical thickness.

sTable 5.8. Statistical significance for all classifiers.

		SB-ROIs	VWGMV	VWCT
Site 1 Chengdu, China	KNN	.003	.001	.002
	LR	.003	.003	.001
	SVM	.003	.004	.013
	DNN	.001	.008	.001
Site 2 London, England	KNN	.083	.891	.200
	LR	.381	.346	.009
	SVM	.757	.207	.450
	DNN	.014	.593	.265
Site 3 Santander A, Spain	KNN	.004	.444	.011
	LR	.021	.002	.013
	SVM	.005	.001	.036
	DNN	.001	.448	.010
Site 4 Santander B, Spain	KNN	.041	.003	.028
	LR	.129	.012	.001
	SVM	.081	.032	.030
	DNN	.001	.014	.002
Site 5 Utrecht, The Netherlands	KNN	.237	.163	.262
	LR	.033	.003	.007
	SVM	.007	.004	.408
	DNN	.108	.010	.008

KNN: k-nearest neighbours; LR: logistic regression; SVM: support vector machines; DNN: deep neural network; SB-ROIS: surfaced-based regional volumes and cortical thickness; VBM: voxel-based morphometry; VBCT: voxel-based cortical thickness.

sTable 5.9. Odds ratio, confidence interval and p-value for the effects of anti-psychotic medication and psychotic symptoms on predicted labels.

		SM-ROIs			VWGMV			VWCT		
		Anti-psychotic medication	Positive symptoms	Negative symptoms	Anti-psychotic medication	Positive symptoms	Negative symptoms	Anti-psychotic medication	Positive symptoms	Negative symptoms
SITE 1	KNN	-	0.99 [0.97-1.09], .656	0.98 [0.97-1.09], .575	-	1.03 [0.97-1.09], .333	1.0 [0.94-1.04], .587	-	1.03 [0.97-1.1], .334	1.0 [0.94-1.05], .768
	LR	-	1.0 [0.97-1.09], .896	1.03 [0.97-1.09], .264	-	0.97 [0.91-1.03], .276	1.0 [0.97-1.07], .516	-	1.01 [0.95-1.07], .783	1.0 [0.99-1.1], .117
	SVM	-	1.0 [0.97-1.09], .896	1.03 [0.97-1.09], .264	-	0.99 [0.93-1.05], .754	1.0 [0.93-1.03], .477	-	0.96 [0.87-1.06], .464	1.0 [0.89-1.08], .676
	DNN	-	0.97 [0.97-1.09], .378	1.03 [0.97-1.09], .351	-	0.99 [0.93-1.05], .724	1.0 [0.95-1.05], .957	-	1.05 [0.98-1.11], .142	1.0 [0.95-1.06], .905
SITE 2	KNN	1.0 [1.0-1.01], .038	1.0 [0.91-1.11], .971	1.0 [0.89-1.05], .447	1.0 [1.0-1.01], .292	0.91 [0.81-1.03], .135	1.0 [0.9-1.07], .641	1.0 [1.0-1.0], .875	1.01 [0.9-1.13], .929	1.0 [0.92-1.1], .864
	LR	1.0 [1.0-1.0], .990	1.02 [0.93-1.12], .694	1.0 [0.9-1.05], .433	1.0 [1.0-1.0], .683	0.99 [0.9-1.09], .887	1.0 [0.94-1.09], .740	1.0 [1.0-1.0], .463	0.93 [0.83-1.05], .262	1.0 [0.94-1.14], .468
	SVM	1.0 [1.0-1.0], .680	0.86 [0.76-0.97], .011	1.0 [0.92-1.09], .889	1.0 [1.0-1.01], .212	0.96 [0.87-1.07], .475	1.0 [0.9-1.06], .535	1.0 [0.9-1.12], .949	1.0 [1.0-1.0], .775	1.0 [0.95-1.1], .673
	DNN	1.0 [1.0-1.0], .956	0.95 [0.85-1.05], .289	1.0 [0.87-1.03], .192	1.0 [1.0-1.01], .275	1.04 [0.94-1.14], .485	1.0 [0.88-1.03], .241	1.0 [1.0-1.0], .344	0.96 [0.86-1.08], .518	1.0 [0.89-1.06], .503
SITE 3	KNN	1.0 [1.0-1.0], .522	0.95 [0.86-1.04], .253	1.0 [0.96-1.11], .456	1.0 [1.0-1.0], .916	1.06 [0.96-1.18], .236	1.0 [0.99-1.16], .092	1.0 [1.0-1.0], .741	0.96 [0.88-1.04], .268	1.0 [0.96-1.09], .515
	LR	1.0 [1.0-1.0], .516	1.01 [0.92-1.11], .799	1.0 [0.91-1.05], .561	1.0 [1.0-1.0], .285	0.97 [0.88-1.07], .548	1.0 [1.0-1.2], .049	1.0 [1.0-1.0], .098	1.0 [0.92-1.08], .942	1.0 [0.94-1.07], .930
	SVM	1.0 [1.0-1.0], .188	1.05 [0.95-1.15], .357	1.0 [0.89-1.04], .320	1.0 [1.0-1.0], .560	0.96 [0.87-1.05], .366	1.0 [0.95-1.12], .416	1.0 [1.0-1.0], .620	0.94 [0.82-1.08], .400	1.0 [0.89-1.1], .797
	DNN	1.0 [1.0-1.0], .277	1.01 [0.92-1.11], .825	1.0 [0.91-1.06], .593	1.0 [1.0-1.0], .926	0.96 [0.88-1.06], .414	1.0 [0.96-1.11], .459	1.0 [1.0-1.0], .817	1.01 [0.93-1.09], .815	1.0 [0.94-1.08], .806
SITE 4	KNN	1.0 [1.0-1.0], .661	0.92 [0.79-1.09], .343	1.0 [0.78-1.02], .086	1.0 [1.0-1.0], .380	1.01 [0.92-1.12], .778	1.0 [0.93-1.11], .697	1.0 [1.0-1.0], .661	1.0 [1.0-1.0], .735	1.0 2 [0.91-1.12], .534
	LR	1.0 [1.0-1.0], .389	1.04 [0.94-1.15], .439	1.0 [0.84-1.0], .050	1.0 [1.0-1.0], .769	1.08 [0.99-1.19], .089	1.0 [0.89-1.04], .328	1.03 [0.95-1.1], .463	1.0 [1.0-1.0], .385	1.0 [1.0-1.0], .638
	SVM	1.0 [1.0-1.0], .584	1.01 [0.92-1.1], .894	1.0 [0.85-0.99], .033	1.0 [1.0-1.0], .126	1.04 [0.94-1.14], .482	1.0 [0.88-1.05], .355	1.0 [1.0-1.0], .884	1.0 [1.0-1.0], .472	0.96 [0.89-1.03], .264
	DNN	1.0 [1.0-1.0], .621	1.04 [0.94-1.15], .493	1.0 [0.85-1.01], .072	1.0 [1.0-1.0], .572	0.95 [0.85-1.06], .356	1.0 [0.82-0.99], .034	1.0 [1.0-1.0], .521	1.02 [0.89-1.14], .573	1.0 [1.0-1.0], .647
SITE 5	KNN	1.0 [0.99-1.0], .207	0.99 [0.89-1.1], .847	1.0 [0.92-1.08], .947	1.0 [1.0-1.0], .375	1.04 [0.93-1.16], .464	1.0 [0.91-1.08], .828	1.0 [1.0-1.0], .489	0.96 [0.84-1.09], .535	1.0 [0.98-1.18], .128
	LR	1.0 [1.0-1.0], .195	0.85 [0.75-0.97], .017	1.0 [0.87-1.04], .267	1.0 [1.0-1.0], .306	1.06 [0.96-1.18], .260	1.0 [0.91-1.06], .615	1.0 [1.0-1.0], .475	0.94 [0.83-1.07], .365	1.0 [0.93-1.1], .836
	SVM	1.0 [1.0-1.0], .313	0.99 [0.89-1.1], .795	1.0 [0.85-1.02], .108	1.0 [0.99-1.0], .105	1.06 [0.94-1.18], .338	1.0 [0.86-1.02], .160	1.0 [1.0-1.0], .400	1.03 [0.91-1.16], .647	1.0 [0.87-1.03], .226
	DNN	1.0 [0.99-1.0], .112	0.98 [0.87-1.1], .706	1.0 [0.81-0.99], .027	1.0 [1.0-1.0], .318	0.96 [0.86-1.06], .408	1.0 [0.98-1.15], .169	1.0 [1.0-1.01], .097	0.95 [0.83-1.08], .443	1.0 [0.82-1.0], .048

OR: odds ratio; CI: confidence interval; KNN: k-nearest neighbour; LR: logistic regression; SVM: support vector machine; DNN: deep neural network.

5.3. sDiscussion

5.3.1. Association between sample size and classification accuracy

Accuracy and sample sizes from existing studies using machine learning and sMRI were extracted as follows:

- up until and including 2013, this information was taken from the latest meta-analysis Kambeitz et al (Kambeitz et al., 2015);

- from 2014 to 2016 this information was taken from the review Arbabshirani et al (2017);

- seven further subsequent studies were identified: Pinaya et al (2016); Salvador et al (2017); Winterburn et al (2017); Xiao et al (2017); Rozycki et al (2018); Dluhoš et al (2017) and de Moura et al (2018). Xiao et al (2017) was excluded as it was a clear outlier (see Figure 5.3A).

Pearson's correlation was used to test for the association between all sample sizes and accuracies. The same studies were used for Figure 5.3A.

5.3.2. Publication bias

Sample size, true positives, false positives, true negatives and false negative scores from each study were extracted from Kambeitz et al (2015). In addition, the main result from the meta-analyses of machine learning-sMRI studies in psychosis (established schizophrenia and FEP combined) was also extracted from the same study. Publication bias was assessed using the same procure as in Kambeitz et al (2015) which in turn was based on recommendations for diagnostic classification studies described in Deeks et al (2005). Briefly, a measure of sample size and effect size were calculated as follows:

- Effective sample size (ESS) was calculated from the patients and control groups sample size using the formula:

$$1/\sqrt{ESS} = \frac{1}{\sqrt{\frac{4n_1 n_2}{n_1 + n_2}}}$$

- lnDOR (diagnostic odds ratio) was calculated using the following formula:

$$DOR = \ln\left(\frac{TF/_{PN}}{FP/_{TN}}\right)$$

The resulting funnel plot was tested for asymmetry through a regression analysis weighted by ESS as implemented in R statistical programming language version 1.1.453 (R Core Team, 2016).

Chapter 6

Using deep learning and structural data to identify first-episode psychosis: a multi-centre mega-analysis

6.1. Introduction

Despite the impressive advances in the understanding of the neurobiological basis of psychiatric disorders in the last decades, there are growing concerns about the reliability and reproducibility of most findings (Anonymus, 2013). Perhaps the most noteworthy source of concern is the high risk of false positives (Button et al., 2013) and heterogeneous findings (IntHout et al., 2015) associated with the small studies that dominate most of the neuroscientific literature. The pressing need for larger samples has led to exceptional efforts (Landhuis, 2017; Poldrack & Gorgolewski, 2014; Smith & Nichols, 2018; Van Horn & Toga, 2014) such as large consortia, including the Human Connectome Project (HCP) (Van Essen et al., 2013), ENIGMA (Bearden & Thompson, 2017), ADNI (Mueller et al., 2005b) and UK Biobank (Sudlow et al., 2015), as well as several neuroimaging data sharing initiatives (Eickhoff et al., 2016; Ferguson et al., 2014; Woo et al., 2017). In psychosis for example, the ENIGMA consortium has led to unprecedented sample sizes in ChSz research, with two recent studies of neuroanatomical abnormalities in a total sample of 9572 (van Erp et al., 2018) and 4568 (van Erp et al., 2016) participants. In the first effort to combine several publicly available datasets of ChSz, Gupta et al. (C. N. Gupta et al., 2015) analysed 784 patients and 936 healthy controls collected from 23 sites. More recently, Rozycki (2018) analysed data from 5 sites totalling 448 healthy controls and 387 ChSz patients.

While such movement is paving the way for more reliable and reproducible findings, there is also a growing demand for clinically translatable research (Borgwardt & Fusar-Poli, 2012). Machine learning is an area of artificial intelligence that promises to meet this demand by being able to learn meaningful patterns from the data and using this information to make predictions about unseen individuals (Hastie, 2009). The promise of this new approach has led to a surge of studies in the last decade across the field of psychiatric neuroimaging, most of which based on neuroanatomical data (Arbabshirani et al., 2017; Wolfers et al., 2015; Woo et al., 2017), including in psychosis (Kambeitz et al., 2015; Orrù et al., 2012; Zarogianni et al., 2013). Nevertheless, the overwhelming majority of studies so far have been conducted in small local samples (Kambeitz et al., 2017, 2015), which have been shown to yield unreliable results (Nieuwenhuis et al., 2012; Schnack & Kahn, 2016; Varoquaux, 2017). By capitalizing on the growing number of neuroimaging consortia and data-sharing initiatives, machine learning mega-studies are now

starting to emerge in the hope of finding more robust and generalizable multivariate biomarkers that can be used for predictions at the level of the individual (e.g. Nunes et al., 2018; Wegmayr, Aitharaju, & Buhmann, 2018; Zhang-James et al., 2019). In most studies, machine learning models are trained and tested by either i) pooling all data together and using a standard CV (e.g. 10-fold CV) – pooled validation; or ii) training the model in all sites but one which is used for testing, until all sites have been used for testing (i.e. leave-one-site-out CV) – cross-site validation. In psychosis, Rozycki et al. (2018) used neuroanatomical data collected from 440 patients diagnosed with ChSz and 501 controls to successfully distinguish the two groups with a promising accuracy of 76% for the pooled data and 75% for the cross-site validation. More recently, Schwarz et al. (2019), was able to discriminate ChSz and controls with an AUC-ROC (area under the receiver operating characteristic curve) of 74% and 71% for the pooled and cross-site validation, respectively, using structural data of 375 patients and 1729 controls. However, given the well-established effects of illness chronicity and antipsychotic medication on brain structure (Bora et al., 2011; Antonio Vita et al., 2015) it is unclear to what extent classification was based on neuroanatomical changes associated with these factors rather than the onset of the illness per se. De Pierrefeu (2018) addressed the issue of chronicity by classifying a smaller sample of 276 ChSz and 330 controls with an accuracy of 72% for the cross-site validation. The same models were subsequently validated in an independent sample of 43 FEP and 90 controls with accuracies of 73%, suggesting some overlap between the features that distinguish ChSz and FEP from controls. To date however, there have been no large-scale machine learning studies trained in a large sample of FEP individuals, which would be more likely to unveil more specific patterns characteristic of the early stages of psychosis.

Although traditional machine learning approaches, such as support vector machine and logistic regression-based models used the studies above, remain popular techniques, an alternative family of methods known as deep learning is gaining considerable attention in the wider research community (Gulshan et al., 2016). Deep learning is a family of representation-learning methods capable of detecting multiple levels of latent representations from the data (LeCun et al., 2015). This is achieved by combining consecutive layers of simple nonlinear transformations that allow the extraction of increasingly abstract features, which may be particularly suitable to model the

subtle, widespread and heterogenous neuroanatomical abnormalities characteristic of the early stages of psychosis (Plis et al., 2014; Schnack, 2017). Preliminary evidence across several psychiatric and neurologic disorders has shown promising results (Durstewitz et al., 2019; Vieira et al., 2017). The aim of this study was to apply a deep learning model to neuroanatomical data to identify individuals at the early stages of psychosis, where the effect to confounding variables is minimal, from controls in a large sample of 958 participants. It is hypothesised that: i) the deep learning model will outperform traditional classifiers and ii) classification will be mostly driven by fronto-temporal brain regions as well as the insula and cingulate.

6.2. Methods

6.2.1. Participants

In this study, all data described in Chapter 2 was combined to create a large multi-centre dataset. The recruitment criteria for each site are reported in section 2.1.2 in Chapter 2. Briefly, patients experiencing their first psychotic episode as defined by either the DSM (APA, 2000) or ICD-10 (Organization World Health, 1992) and controls from the same geographical area were recruited as part of five independent studies carried out in four sites: China (Gong et al., 2015), England (Di Forti et al., 2009), Spain (Pelayo-Terán et al., 2008) and The Netherlands (Korver et al., 2012). At each site, data was match between FEP and HC with respect to sex and age (+/- 5years) to mitigate unwanted effects from these demographic variables (for more details see section 5.1.1.2 in the supplementary materials in Chapter 5). The final demographic and clinical characteristics for each site and combined data are reported in Table 6.2.

6.2.2. Magnetic resonance imaging

6.2.2.1. Acquisition

At all five sites, structural magnetic resonance imaging (sMRI) data was acquired using a T1-weighted protocol with a SPGR sequence. The details of the image acquisition sequence varied between scanners, as reported in Chapter 2, section 2.3.2.

6.2.2.2. MRI preprocessing

The T1-weighted images were preprocessed using the FreeSurfer image analysis package

(version 5.3.0, http://surfer.nmr.mgh.harvard.edu/) (Fischl, 2012). FreeSurfer is a widely used fully automated suite of tools capable of estimating a wide range of surface-base morphometric measures including the volume of subcortical structures as well as several cortical metrics such as cortical thickness and volume. In this study, the volume of 33 subcortical brain regions as well as the volume and thickness of 68 cortical regions, as defined by the Desikan-Killiany atlas (Desikan et al., 2006), were extracted using the 'recon-all' command. For a complete list of the brain regions included, please refer to Table 2.1 in Chapter 2. The details of the subcortical and cortical preprocessing stages are described in section 2.2.3.2 in Chapter 2. Briefly, each image was corrected for intensity normalization, followed by the removal of non-brain tissue, segmentation of the subcortical WM and deep GM volumetric structures, tessellation of GM and WM boundaries, automated topology correction, surface deformation, registration to a spherical atlas and parcellation based on a spherical in-built atlas to extract morphometric measurements for specific cortical regions (Dale et al., 1999; Fischl et al., 1999).

6.2.3. Deep neural network

6.2.3.1. Model specification

Deep neural networks (DNNs) are a deep learning general-purpose network characterized by its distinctive multi-layered and fully-connected architecture. Its layer-wise structure allows for successive nonlinear transformations of the input data, such that the network learns increasingly abstract features. These are then used to perform a task, for example distinguish between two groups (Bengio et al., 2015). For a detailed description of the structure and training procedure involved in DNNs see section 2.3.2.4.2 in Chapter 2.

In this study, the volumes of 33 subcortical brain regions as well as the thicknesses and volumes of 34 cortical regions form each hemisphere were concatenated to form a 1-dimensional (1D) vector totalling 169 features (Figure 6.1).

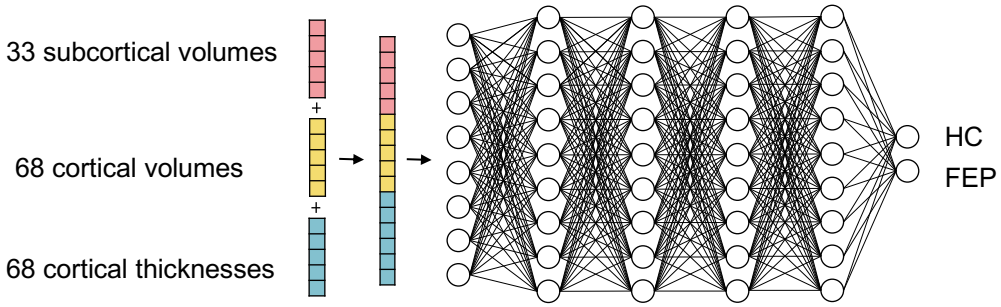


Figure 6.1. DNN structure**.** The volumes and thicknesses of subcortical and cortical brain regions were concatenated into a 1D vector and used as input features for a DNN.

The weights of the DNN were initialized via Glorot (also known as Xavier) initialization (normal distribution) (Glorot & Bengio, 2010) and subsequently adjusted using backpropagation and a gradient descent-based optimizer with a mini-batch size of 128 training examples. ReLu was chosen as the activation function at each neuron and in the output layer, a softmax function was used to perform the binary classification (FEP or HC). As in Chapter 5, the optimal number of layers, number of neurons in each layer, optimizer, learning rate, learning rate decay, epoch, as well as L2 regularizer (Krogh & Hertz, 1992) and dropout (Srivastava et al., 2014) were determined using nested cross-validation (CV) (see section 2.4.2 in this Chapter).

The motivation for the use of SB-ROIs in combination with DNNs was two-fold. First, the dimensionality of this type of data is considerably lower compared to the standard whole-brain voxel-level data, for example. In the context of deep learning, this has important implications in terms of overfitting and computational resources. As a nonlinear complex approach, i.e. with many parameters to estimate, deep learning is particularly prone to overfitting compared to traditional machine learning methods. Therefore, a reduced number of input features will decrease the likelihood of overfitting by reducing model complexity as well as alleviate computational requirements. Although the dimensionality of voxel-level data could have been addressed with the use of more sophisticated models such as autoencoders, CNNs or a combination of the two, this would become prohibitively expensive for amount of data to be analysed and the computational resources available. In addition, the combination of SB-ROIs and DNN also

allowed for the automatic, as opposed to manual, tuning of hyperparameters, therefore reducing the risk of overfitting. Second, the results from the site-level analysis in Chapter 5 showed a superior, albeit modest, performance of SB-ROIs over the other types of features included in this work, namely VWGMB and VWCT.

6.2.3.2. Model training

The DNN was trained using two validation schemes: stratified 10-fold CV (pooled validation) and leave-one-site-out (LOSO) CV (cross-site validation). Both validation types included a stratified nested CV for hyperparameter tuning to minimize bias during model selection, consistent with recommended practice (Varma & Simon, 2006). For the pooled validation, this first involved partitioning the total data into 10 parts with the same proportion of HC and FEP. Nine parts were combined to create a training set and the remaining was used as the test set; this split training/test defines the first iteration of the outer CV. For the cross-site validation, four sites are used for training, and one site is used for testing. For the remaining process, the procedure is the same for both validation types. The training set is further divided into 10 parts, with the same proportion of HC and FEP. Nine parts are combined to create a new training set, and the part left-out was used as the validation set. With these sets defined, a random selection of hypermeters from an a priori defined search space (Table 6.1) is chosen to build a DNN with, for example, number of layers=3, number of neurons in each layer=50, learning rate=0.01, learning rate decay=10^{-4}, epochs=100, optimizer=SGD, momentum=0.9, L2 norm=10^{-3} and dropout rate=0.5. This DNN is then trained on the new training set and its balanced accuracy is estimated in the validation set. This process was repeated 10 times using the same combination of hyperparameters, each time with one of the 10 possible validation sets. The cross-validated mean balanced accuracy for that particular combination of hyperparameters was then estimated. This entire process was repeated 500 times, each time with a different random combination of parameters (different number of layers, number of neurons, different activation function, etc). This resulted in a total of 500 cross-validated mean balanced accuracies. The combination of hyperparameters that yielded the best performance was then used to train a DNN again in the whole training data set as defined by the outer CV and test it in the test set. This process is done iteratively 10 and 5 times for the pooled and cross-site validations, respectively, each time with a different training and test sets as defined

by the outer CV. The final balanced accuracies are averaged to estimate the model's final performance.

Table 6.1. Search space for each DNN hyperparameter.

Parameter	Values
Number of layers	2, 3, 4, 5
Number of units	10, 20, 50, 75, 100, 150
Learning rate	0.001, 0.005, 0.01, 0.1, 0.2
Learning rate decay	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}
Epochs	50, 100, 150
Optimizer	Stochastic gradient descent (SGD), Adam
Momentum	0.99, 0.9, 0.95
L2 norm	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}
Dropout rate	0.2, 0.5, 0.7

6.2.4. Traditional machine learning algorithms

Logistic regression and support vector machine, two popular and well-established machine learning techniques in psychiatric neuroimaging, were also used for comparison.

6.2.4.1. Logistic regression

A logistic regression (LR) is a simple yet powerful supervised algorithm that aims to find the optimal linear combination of the input features and outputs the probability of the input data belonging to a default class (e.g. HC), which can then be converted to a binary prediction. In this study, LR was implemented via elastic net, a regularized regression that combines the regularizations L1 norm and L2 norm. The former retains all variables and minimizes the impact of irrelevant features, therefore reducing the model's dependency on a specific group of features from the training set. The latter on the other hand, discards unimportant variables reducing the overall model complexity (H. Zou & Hastie, 2005). The optimal relative contribution of each penalty was determined by tuning the ratio between the two via grid search. The value for this ratio was chosen from eleven possible values between 0 and 1 with increments of 0.1 via nested stratified 10-fold CV, where lower values indicate a larger contribution from L2 relative to L1 while values closer to 1 indicate the opposite.

6.2.4.2. Support vector machine

Support vector machine (SVM) is a popular supervised machine learning technique that aims to classify data points by maximising the margin between classes in a high-dimensional space (Pereira et al., 2009; Vapnik, 1995). This is achieved by projecting the data into a feature space using a similarity function, known as a kernel. Here, the algorithm is trained to find the optimal separating hyperplane by maximising the margin between the examples lying closest to the separating plane (and hence the most difficult to classify), known as the support vectors. This particular hyperplane is learned from the training data and subsequently used as a decision boundary where observations in the test set falling on either side of the hyperplane are assigned to either class (Noble, 2006). In this study, a linear kernel was chosen to contrast with the characteristic nonlinear approach of deep learning. The soft margin (C) parameter, that controls the trade-off between having zero training errors and allowing misclassifications, was tuned from a possible range of values (2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}) using grid search via nested stratified 10-fold CV.

6.2.5. Model performance

The final performance for each classifier was assessed by estimating the average balanced accuracy, sensitivity and specificity across the 10 and 5 iterations of the pooled and cross-site CVs, respectively. The balanced accuracy of each classifier was tested for significance using permutation testing. This consisted of randomly assigning participants to one of the classes (FEP/HC) and run the same pooled/cross-site CV model. This procedure was repeated 1000 times. This resulted in a distribution of accuracies reflecting the null hypothesis that the classifier did not exceed chance. The number of times the classifier's performance was greater than or equal to the true accuracy was divided by 1000 to determine a *p*-value. A *p*-value lower than 0.05 was considered statically significant.

6.2.6. Effect of scanner

The effect of scanner on each individual feature was mitigated using a linear regression model as implemented by the OLS function from the statsmodels library (version 0.10.1) for Python (Seabold & Perktold, 2010). Five hot-one-encoded variables, one coding each site, were entered as the independent variables in a series of linear regression models, each one predicting one of

the input features. The resulting residuals for each feature were standardized by removing the mean and scaling to unit variance using the StandardScaler from the preprocessing module of the sklearn library (v0.20) (Pedregosa et al., 2011). The standardized residuals subsequently used as input features for each one of the classifiers. Importantly, this procedure was implemented within the CV framework to avoid knowledge-leakage between training and test sets. In practice, this involved fitting each linear regression to the training data and use the resulting regression parameters to estimate the residuals in the training and test sets. Likewise, standardization consisted in estimating the mean and standard-deviation for each feature in the training set and used to standardize the training and test sets.

6.2.7. Most contributing brain regions

The most contributing features to discriminate between FEP and HC for the two DNN models – pooled and cross-site validation – were identified with the function SmoothGrad (Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017) as implemented by the iNNvestigate library (Alber et al., 2018). Briefly, SmoothGrad works by first adding Gaussian noise to several copies of the input data. Each copy is then put through the trained model and a saliency map is generated from the network's gradients. This results in several saliency maps that are then average to estimate a final smoothed saliency map. Smoothgrad takes two parameters: noise level or standard deviation of the Gaussian perturbations, and n, the number of samples to average over. Here we use the default parameters, standard deviation of 0.1, and 64 copies of in the input. The ranked lists of features from each CV modality were compared using Kendall's Tau-b, a rank correlation coefficient that measures the degree of similarity between two rankings.

6.2.8. Experiments

All experiments were done in Python (version 3.6). DNNs were implemented with Tensorflow v.1.4 (Abadi, Agarwal, et al., 2016) and Keras v.2.1 (https://keras.io/) libraries. Both LR and SVM were implemented using the Scikit‐learn library (version 0.19.2; Pedregosa et al., 2011). The same random seed was used across all classifiers to ensure the starting weights and the split of participants during the cross-validation was equivalent in every experiment.

6.3. Results

6.3.1. Demographic and clinical characteristics

No statistically significant differences were identified between patients and controls for age, sex at each site and in the combined dataset (Table 6.2).

6.3.2. Pooled validation

In the pooled validation analysis, the DNN model was able to classify patients and controls with 65.4% balanced accuracy, 63.3% sensitivity and 67.5% specificity. Although a modest performance, it was able to distinguish the two groups with a higher performance compared to LR and SVM with a balanced accuracy of 58.1% and 61.6%, respectively (Table 6.3).

Table 6.3. Balanced Accuracy, sensitivity and specificity for the pooled validation.

	Balanced Accuracy	Sensitivity	Specificity
LR	61.0±5.2***	65.2±13.7	56.8±7.9
SVM	61.6±4.1***	64.6±6.8	58.6±4.7
DNN	65.4±3.9***	63.3±7.1	67.5±5.3

LR: logistic regression; SVM: support vector machine; DNN: deep neural network; ***p<.001

6.3.3. Cross-site validation

The classification accuracy when all but one of the sites were used for training yielded a balanced accuracy of 59.3% for the DNN and 56.8% and 58.0% for the LR and SVM models, respectively.

Table 6.4. Balanced accuracy, sensitivity and specificity for LOSO CV.

	Balanced Accuracy	Sensitivity	Specificity
LR	56.8±2.4***	60.4±18.5	53.3±18.8
SVM	58.0±3.9***	72.4±11.7	43.5±17.0
DNN	59.3±3.1***	63.5±8.6	55.1±6.1

LR:logistic regression; SVM: support vector machine; DNN: deep neural network; ***p<.001

The optimized hyperparameters for all models is shown in sTables 6.1 and sTable 6.2 in the supplementary materials.

Table 6.2. Demographic and clinical characteristics for FEP and HC for each site and combined data.

		Chengdu, China (N=224)		London, England (N=142)		Santander A, Spain (N=220)		Santander B, Spain (N=210)		Utrecht, The Netherlands (N=162)		Combined data (N=958)	
		HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP
N		112	112	71	71	110	110	70	140	81	81	444	514
	M	51 (46)	51 (46)	36 (51)	36 (51)	68 (62)	68 (62)	45 (64)	90 (64)	64 (79)	64 (79)	264 (59)	309 (60)
Sex (%)	F	61 (54)	61 (54)	35 (49)	35 (49)	42 (38)	42 (38)	25 (46)	50 (46)	17 (21)	17 (21)	180 (41)	205 (40)
		χ^2=ns		χ^2=ns		χ^2=ns		χ^2=ns		χ^2=ns		χ^2=ns	
Age M(SD)		27.2 (7.3)	25.7 (8.1)	26.8 (7.1)	26.4 (6.2)	29.7 (7.8)	28.5 (8.6)	27.3 (7.5)	28.3 (7.6)	26.9 (8.0)	25.2 (5.9)	27.6 (7.5)	26.8 (7.3)
		t=ns		t=ns		t=ns		t=ns		t=ns		t=ns	
TIV (L) M(SD)		1.5 (0.1)	1.5 (0.2)	1.5 (0.2)	1.5 (0.2)	1.5 (0.1)	1.4 (0.2)	1.5 (0.1)	1.5 (0.1)	1.6 (0.1)	1.5 (0.2)	-	-
		t=ns		*t*=ns		*t*=ns		*t*=ns		*t*=ns			
Positive symptoms M(SD)		-	24.6 (6.6)[a]	-	13.9 (5.5)[a]	-	14.7 (4.6)[b]	-	14.4 (4.1)[b]	-	15.9 (6.3)[a]	-	-
Negative symptoms M(SD)		-	18.2 (7.7)[a]	-	16.0 (6.0)[a]	-	6.3 (4.6)[c]	-	6.1 (5.0)[d]	-	16.2 (6.9)[a]	-	-
Duration of illness (years) Med (IQR)		-	0.3 (1.1)	-	1.1 (0.3)	-	0.3 (0.7)	-	0.3 (0.9)	-	0.6 (1.0)	-	-

TIV: total intra-cranial volume; L: litres; M: male; F: female; FEP: first episode psychosis; HC: healthy controls. [a]PANSS: Positive and Negative Symptoms Scale; [b]SAPS: Scale for the Assessment of Negative Symptoms; [c]SANS: Scale for the Assessment of Negative Symptoms ns: $p<0.05$

6.3.4. Most contributing features

The top features driving the predictions of the DNN in the pooled and cross-site validation are displayed in Figure 6.2. For a complete list of regions and their respective weights and rankings, see sTable 6.1 in the supplementary materials.

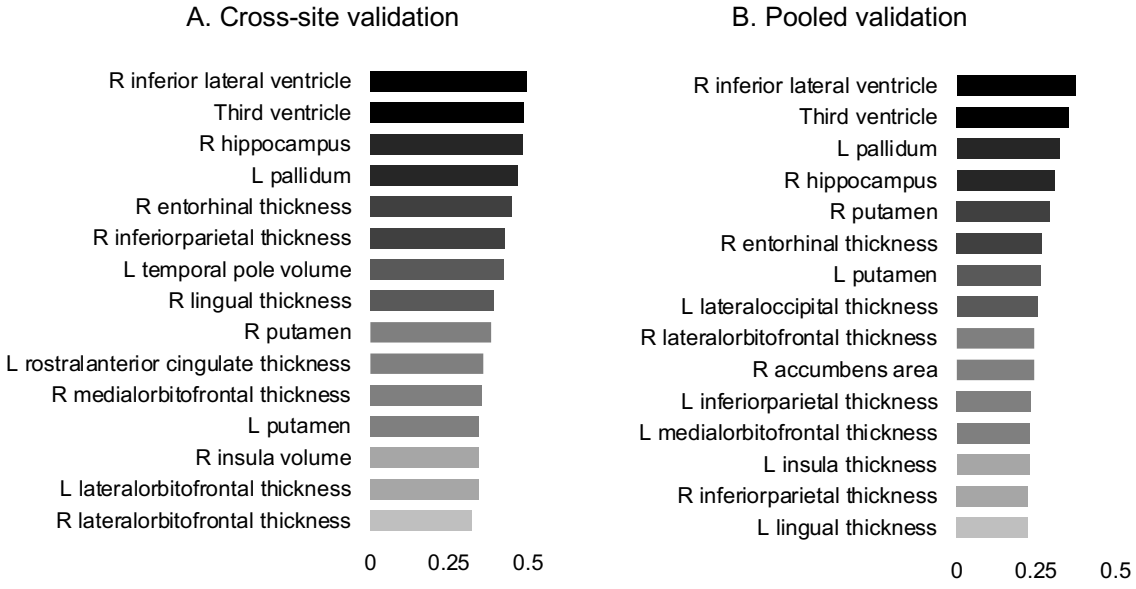


Figure 6.2. Top 15 regions with the highest weights. L: left, R: right.

It can be seen that the right inferior lateral ventricle, third ventricle, right hippocampus and left pallidum were the four brain regions with the largest weights in both validation schemes. Other common regions, albeit some in different hemispheres, included the putamen, insula and entorhinal cortex as well as the lingual, inferior parietal, lateral and medial orbitofrontal gyri. The correlation between the rankings of the regions from each validation schemes was moderate when all regions were considered (r_τ=.31, p<.001) and high for top five brain regions (r_τ=.73, p<.05). The coefficients for the LR and SVM as well as their respective overall weight-based correlations are reported in the supplementary materials in sTable 6.4 and sTable 6.5., respectively.

6.4. Discussion

The increasing efforts to overcome the limitations of small samples combined with the demands for translatable research is shifting psychiatric neuroimaging research away from local group-level

findings towards large-scale machine learning studies in order to find reliable multivariate markers that can be used in clinical practice. The aim of this study was to use a promising machine learning method, known as deep learning, to discriminate individuals with a recent first psychotic episode from controls at the individual level, based on a large sample of neuroanatomical data.

Overall, the accuracies obtained from the pooled data were much lower than the ones reported by most single-site studies (Kambeitz et al., 2015). At least two reasons may explain this discrepancy. First, smaller studies tend to yield unstable performances (Nieuwenhuis et al., 2012; Varoquaux, 2017) which, in combination with less-than-rigours methods (Arbabshirani et al., 2017; Wolfers et al., 2015) and publication bias speculated elsewhere (Schnack & Kahn, 2016) and shown in Chapter 5, may have contributed to the over-representation of inflated results from small local single-site studies. Second, finding a pattern of shared abnormalities in patients relative to controls is likely to be easier in smaller homogeneous samples recruited with stringent inclusion criteria compared larger studies with loosen criteria that result in more heterogenous samples. In a multi-site study made up of several relatively large single-site samples such as the present study, this is further exacerbated by the use of different diagnostic criteria, inclusion criteria, assessment protocols, including different scanners and acquisition parameters across the different sites (Schnack, 2017). This results in a trade-off between homogeneous small-sample and heterogeneous large-sample studies: while the former tends to perform well at the cost of lower generalizability, the latter are more likely to yield lower accuracies but with better generalizability since it is expected that whatever the pattern identified, it will be more representative of the alterations in that diagnostic group (Schnack & Kahn, 2016). Building predictive models that yield potentially high accuracies at the expense of generalizability has, of course, limited translational potential. Therefore, despite less encouraging, the findings from this study are likely to be more reliable than those reported in initial small studies. This inversed relationship between performance and sample size was initially observed when single-site samples increased from a few dozen to a few hundred (Schnack & Kahn, 2016; Wolfers et al., 2015). During the last two years, as samples continue to increase even further to several hundred or even a few thousand participants, this trend is becoming more apparent, possibly due to the added heterogeneity from combining several independent datasets. For example, Nunes et al.

(2018) reported a classification accuracy of 65.2% in a sample of 853 individuals diagnosed with bipolar disorder and 2167 controls, while some of the single-site accuracies from sites included in the same study were as high as 81.1%. Faraone et al. (2019) were able to distinguish 1393 patients with attention-deficit/hyperactivity disorder from 1320 controls with an AUC-ROC of 67%, although findings at the single-site level tend to be much higher (Wolfers et al., 2015). The same trend is observed in ChSz. Rozycki (2018) and Schwarz (Schwarz et al., 2019) reported an accuracy and AUC-ROC of 76% and 74% in a sample of 941 and 2014 participants, respectively, which, although encouraging for such large samples, are lower than many smaller studies (Kambeitz et al., 2015; Wolfers et al., 2015; Zarogianni et al., 2013).

As hypothesised, the DNN was able to discriminate the two groups with better performance than the traditional approaches, suggesting that modelling complex nonlinear relationships between brain regions may be useful to identify the neuroanatomical abnormalities in the early stages of psychosis. This is consistent with the initial evidence shown in Chapter 4, that deep learning tends to perform similar or better than traditional machine learning. Nevertheless, when the DNN was trained in a group of sites and tested in an independent site (cross-site validation), performance dropped to 59.3%; a similar accuracy compared to that of the traditional approaches. The drop in performance from the pooled to the cross-site validation is not surprising, as the latter is a more demanding test of generalizability (Woo et al., 2017). Similar lower cross-site accuracies were also found by Nunes (2018) with 58.7% in bipolar disorder and in FEP with 62.0% in a study with 480 participants (Dluhoš et al., 2017). Nevertheless, compared to the traditional approaches, the DNN suffered a larger decrease in accuracy. Therefore, the possibility that the superior performance in the pooled data may have steamed from overfitting cannot be ruled out, despite the use of a large sample, relatively low dimensional data and regularization strategies. Based on the above and in the absence of other similarly large studies, we speculate that a reliable classification accuracy of FEP based on sMRI data may be around 60%. The performance of recent large-scale studies in ChSz seem to be converging around 70% (de Pierrefeu et al., 2018; Rozycki et al., 2018; Schwarz et al., 2019), as predicted in a ´accuracy-sample size' model based on ChSz and FEP (albeit in much smaller number) sMRI studies published up until 2016 (Schnack & Kahn, 2016). However, identifying FEP is more challenging than ChSz due to the more subtle

effects (Egerton et al., 2011) combined with the increased heterogeneity characteristic with the early stages of psychosis, and it is therefore reasonable to expect that large-scale FEP studies will converge at an performance lower than ChSz.

The brain regions driving both the pooled and cross-site validation analysis were highly consistent, suggesting a reproducible signature of the main neuroanatomical abnormalities at the individual level. The main common regions comprised the third and inferior lateral ventricles, basal ganglia (putamen and pallidum), temporal regions such as the hippocampus, entorhinal cortex and lingual gyrus, as well as the inferior parietal gyrus, orbitofrontal regions and the insula. These regions have been repeatedly implicated at group-level in both established (C. N. Gupta et al., 2015; van Erp et al., 2016, 2018) and first-episode (X. Gao et al., 2018; Shah et al., 2017) psychosis. Critically, similar patterns of anatomical changes have also been reported in large multivariate studies in ChSz (de Pierrefeu et al., 2018) and FEP (Dluhoš et al., 2017).

6.5. Conclusion and future directions

To our knowledge this is the largest machine learning study in the early stages of psychosis to date. Although less encouraging, the results presented in this study provide a step towards recent calls for a new generation of machine learning applications that favours reliable and generalizable findings from large-scale studies (Schnack & Kahn, 2016; Woo et al., 2017). Furthermore, we also found a reproducible signature of the main neuroanatomical abnormalities driving the distinction between FEP and controls, consistent with the literature. Future studies should include other diagnosis to assess the specificity of these models and their neuroanatomical signatures. For example, the putamen, identified here as an important region for classification, has been identified as a transdiagnostic marker for psychiatric illness (Gong et al., 2018). Importantly, future studies could also investigate whether a non-linear SVM would be able to match or outperform the DNN model, since it is possible that a simpler model than a DNN could reach the same or better result. Critically, results from this study support the premise that neuroanatomical data alone will not be able to identify FEP with the necessary performance for clinical translation. There is initial evidence showing the superiority of functional MRI over sMRI data in diagnostic classification in psychosis (Kambeitz et al., 2015). Future large-scale studies should address the

reliability of such findings, or better yet, combine both modalities to maximise performance using advance data fusion methods (Calhoun & Sui, 2016). In addition, future studies could also leverage on recent sophisticated approaches such as domain adaptation as a potential solution to mitigate inter-scanner differences for cross-site models. Put simply, domain adaptation techniques address the assumption of most machine learning applications that the training and future data must be in the same feature space and have the same distribution, by allowing models to 'transfer knowledge' from one domain (source domain) to another (target domain) (Pan & Yang, 2010). This is a promising approach for neuroimaging as it would allow, for example, to train a model in one scanner and adapt it to another scanner; thus avoiding having to train a model from scratch for each scanner every time. Finally, while developing models to identify the first manifestation of a psychotic disorder is an important endeavour, the main challenge in clinical decision-making remains predicting conversion to illness, disease progression and treatment response. Therefore, similar mega-analytic efforts now emerging for diagnostic classification studies should be extended to longitudinal data to meet the demands for translational research.

Chapter 6 supplementary materials

sTable 6.1. Optimized hyperparameters for each fold of the two DNN models (10-fold CV and LOSO CV). Each cell contains the optimized number of layers, number of units, learning rate, learning rate decay, optimizer, momentum (if optimizer is SGD), L2 norm and dropout rate.

	10-fold CV	LOSO CV
Fold 1	4, 150, .001, .01, SGD, .99, 0, .5	3, 50, .01, .01, SGD, .99, 10^{-5}, .5
Fold 2	4, 75, .005, .001, SGD, .9, 10^{-5}, .5	3, 150, .005, .001, Adam, 10^{-5}, .5
Fold 3	3, 50, .001, .01, Adam, 0, .5	5, 100, .005, .01, Adam, 10^{-5}, .2
Fold 4	4, 75, .001, 10^{-5}, SGD, .9, 10^{-4}, 0.5	5, 100, .001, .001, Adam, 10^{-3}, .5
Fold 5	5, 100, .001, .001. Adam, 0, .5	5, 150, .001, 10^{-4}, Adam, .01, .5
Fold 6	4, 100, .001, 10^{-5}, SGD, .9, 10^{-3}, .5	-
Fold 7	3, 75, .005, .01, Adam, 0, .7	-
Fold 8	5, 100, .001, .001, Adam, 0, .5	-
Fold 9	5, 150, .01, .01, SGD, .95, 10^{-3}, .5	-
Fold 10	4, 150, .001, 10^{-5}, SGD, .99, 10^{-5}, .5	-

sTable 6.2. Optimized hyperparameters for each fold of the two regularized logistic regression and SVM models (10-fold CV and LOSO CV). Each cell contains the optimized value for the L1/L2 ratio and C hyperparameters for the regularized logistic regression and SVM models, respectively.

	10-fold CV		LOSO CV	
	Regularized LR	SVM	Regularized LR	SVM
Fold 1	0.4	0.03125	0.1	0.03125
Fold 2	0.8	0.5	1	0.03125
Fold 3	0.4	2	0.2	0.5
Fold 4	0.2	0.03125	0.5	2
Fold 5	0.7	0.03125	0.1	0.5
Fold 6	0.6	0.03125	-	-
Fold 7	0.5	0.5	-	-
Fold 8	0.2	0.03125	-	-
Fold 9	0.9	0.03125	-	-
Fold 10	0.7	2	-	-

sTable 6.3. Coefficients and ranking for each feature for the pooled and cross-site validations.

	Cross-site validation		Pooled validation	
	Coefficients	Rank	Coefficients	Rank
Right_Inf_Lat_Vent	0.50	1	0.37	1
3rd_Ventricle	0.49	2	0.35	2
Left_Pallidum	0.47	4	0.33	3
Right_Hippocampus	0.48	3	0.31	4
Right_Putamen	0.38	9	0.29	5
rh_entorhinal_thickness	0.45	5	0.27	6
Left_Putamen	0.35	12	0.26	7
lh_lateraloccipital_thickness	0.10	118	0.25	8
rh_lateralorbitofrontal_thickness	0.33	15	0.24	9
Right_Accumbens_area	0.12	108	0.24	10
lh_inferiorparietal_thickness	0.14	92	0.23	11
lh_medialorbitofrontal_thickness	0.20	47	0.23	12
lh_insula_thickness	0.30	20	0.23	13
rh_inferiorparietal_thickness	0.43	6	0.23	14
lh_lingual_thickness	0.22	41	0.23	15
rh_rostralanteriorcingulate_volume	0.25	30	0.23	16
lh_transversetemporal_thickness	0.27	25	0.22	17
Right_Amygdala	0.24	31	0.22	18
lh_temporalpole_volume	0.42	7	0.22	19
rh_cuneus_thickness	0.18	56	0.22	20
CC_Mid_Anterior	0.07	161	0.21	21
rh_medialorbitofrontal_thickness	0.35	11	0.21	22
Left_Inf_Lat_Vent	0.17	69	0.21	23
rh_medialorbitofrontal_volume	0.17	70	0.21	24
rh_transversetemporal_thickness	0.18	55	0.20	25
Left_Amygdala	0.08	152	0.20	26
rh_temporalpole_volume	0.10	121	0.20	27
CC_Central	0.14	88	0.20	28
rh_middletemporal_thickness	0.17	71	0.19	29
rh_caudalanteriorcingulate_volume	0.32	16	0.19	30
lh_paracentral_thickness	0.17	66	0.19	31
th_Ventricle	0.05	168	0.19	32
lh_parstriangularis_volume	0.32	17	0.19	33
rh_pericalcarine_thickness	0.29	22	0.19	34
lh_transversetemporal_volume	0.22	39	0.18	35
rh_parahippocampal_thickness	0.26	26	0.18	36
rh_lingual_thickness	0.39	8	0.18	37
rh_insula_volume	0.34	13	0.18	38
lh_parahippocampal_thickness	0.30	21	0.18	39
lh_pericalcarine_volume	0.11	117	0.18	40
lh_superiorfrontal_thickness	0.27	24	0.18	41
lh_cuneus_volume	0.16	80	0.18	42
rh_temporalpole_thickness	0.17	62	0.18	43
CC_Posterior	0.08	145	0.17	44
rh_isthmuscingulate_volume	0.17	61	0.17	45
lh_precuneus_volume	0.12	106	0.17	46
Right_Cerebellum_Cortex	0.14	91	0.17	47
lh_inferiortemporal_thickness	0.22	43	0.17	48
lh_paracentral_volume	0.09	144	0.16	49
lh_rostralanteriorcingulate_thickness	0.36	10	0.16	50
Left_Lateral_Ventricle	0.09	136	0.16	51
rh_superiortemporal_volume	0.21	45	0.16	52
rh_precuneus_volume	0.09	131	0.16	53
lh_rostralanteriorcingulate_volume	0.08	150	0.16	54
lh_isthmuscingulate_volume	0.13	94	0.16	55
rh_lateralorbitofrontal_volume	0.17	67	0.16	56
lh_bankssts_thickness	0.11	116	0.16	57
rh_rostralanteriorcingulate_thickness	0.13	95	0.16	58
lh_lateralorbitofrontal_volume	0.16	76	0.16	59
lh_precuneus_thickness	0.24	32	0.16	60
rh_superiorfrontal_thickness	0.10	127	0.16	61
lh_frontalpole_thickness	0.19	52	0.16	62
CC_Anterior	0.10	128	0.15	63

lh_entorhinal_thickness	0.15	83	0.15	64
lh_entorhinal_volume	0.22	37	0.15	65
lh_parahippocampal_volume	0.09	134	0.15	66
rh_postcentral_volume	0.17	64	0.15	67
rh_postcentral_thickness	0.23	35	0.15	68
rh_bankssts_volume	0.26	27	0.15	69
lh_insula_volume	0.20	49	0.15	70
rh_parstriangularis_volume	0.16	78	0.15	71
rh_inferiortemporal_volume	0.08	157	0.15	72
rh_parsopercularis_thickness	0.15	86	0.15	73
Right_Pallidum	0.10	130	0.15	74
rh_lateraloccipital_volume	0.15	84	0.15	75
lh_parsopercularis_volume	0.23	34	0.15	76
lh_posteriorcingulate_volume	0.10	124	0.15	77
lh_pericalcarine_thickness	0.08	155	0.14	78
lh_middletemporal_volume	0.11	115	0.14	79
Left_Cerebellum_White_Matter	0.23	36	0.14	80
lh_bankssts_volume	0.06	166	0.14	81
rh_transversetemporal_volume	0.16	81	0.14	82
Left_Hippocampus	0.29	23	0.14	83
CC_Mid_Posterior	0.17	65	0.14	84
rh_parsorbitalis_thickness	0.10	123	0.14	85
lh_lateralorbitofrontal_thickness	0.34	14	0.14	86
rh_middletemporal_volume	0.12	102	0.14	87
rh_lateraloccipital_thickness	0.18	57	0.14	88
rh_supramarginal_volume	0.16	77	0.14	89
rh_precentral_thickness	0.18	58	0.14	90
rh_parsopercularis_volume	0.11	114	0.14	91
Brain_Stem	0.13	97	0.14	92
rh_caudalanteriorcingulate_thickness	0.16	79	0.14	93
lh_parsorbitalis_thickness	0.09	137	0.14	94
lh_middletemporal_thickness	0.22	42	0.14	95
lh_medialorbitofrontal_volume	0.12	110	0.14	96
rh_cuneus_volume	0.14	90	0.14	97
Right_Lateral_Ventricle	0.25	28	0.14	98
lh_fusiform_thickness	0.10	120	0.13	99
Left_Cerebellum_Cortex	0.21	46	0.13	100
lh_postcentral_volume	0.15	85	0.13	101
Left_VentralDC	0.09	142	0.13	102
lh_inferiorparietal_volume	0.21	44	0.13	103
lh_frontalpole_volume	0.10	126	0.13	104
rh_superiortemporal_thickness	0.17	63	0.13	105
rh_superiorfrontal_volume	0.13	93	0.13	106
rh_fusiform_volume	0.08	154	0.13	107
lh_caudalanteriorcingulate_volume	0.09	135	0.13	108
rh_fusiform_thickness	0.08	148	0.13	109
lh_supramarginal_thickness	0.12	103	0.13	110
lh_lateraloccipital_volume	0.09	141	0.13	111
rh_caudalmiddlefrontal_volume	0.24	33	0.13	112
lh_cuneus_thickness	0.12	107	0.13	113
rh_caudalmiddlefrontal_thickness	0.18	54	0.12	114
rh_precentral_volume	0.16	72	0.12	115
lh_posteriorcingulate_thickness	0.25	29	0.12	116
rh_superiorparietal_volume	0.11	113	0.12	117
lh_superiorparietal_thickness	0.12	109	0.12	118
Right_Thalamus_Proper	0.07	160	0.12	119
rh_lingual_volume	0.22	40	0.12	120
rh_parahippocampal_volume	0.12	101	0.12	121
lh_superiortemporal_volume	0.09	143	0.12	122
lh_superiorparietal_volume	0.08	151	0.12	123
CSF	0.09	140	0.12	124
Right_VentralDC	0.11	111	0.12	125
lh_postcentral_thickness	0.07	159	0.12	126
lh_superiorfrontal_volume	0.09	133	0.12	127
rh_supramarginal_thickness	0.22	38	0.12	128
Left_Thalamus_Proper	0.31	19	0.12	129
lh_fusiform_volume	0.06	164	0.12	130
rh_parstriangularis_thickness	0.10	122	0.11	131
lh_isthmuscingulate_thickness	0.12	104	0.11	132

rh_superiorparietal_thickness	0.13	99	0.11	133
Right_Cerebellum_White_Matter	0.20	48	0.11	134
rh_frontalpole_thickness	0.09	139	0.11	135
rh_posteriorcingulate_thickness	0.14	89	0.11	136
rh_inferiortemporal_thickness	0.07	163	0.11	137
lh_parsopercularis_thickness	0.18	60	0.11	138
lh_lingual_volume	0.15	87	0.11	139
rh_paracentral_volume	0.10	125	0.11	140
rh_isthmuscingulate_thickness	0.11	112	0.11	141
lh_superiortemporal_thickness	0.19	50	0.11	142
rh_insula_thickness	0.09	138	0.11	143
rh_frontalpole_volume	0.06	165	0.11	144
rh_rostralmiddlefrontal_volume	0.07	158	0.11	145
Left_Accumbens_area	0.05	167	0.10	146
rh_posteriorcingulate_volume	0.16	74	0.10	147
rh_inferiorparietal_volume	0.08	146	0.10	148
lh_inferiortemporal_volume	0.19	51	0.10	149
rh_paracentral_thickness	0.10	129	0.10	150
rh_entorhinal_volume	0.09	132	0.10	151
rh_bankssts_thickness	0.16	75	0.10	152
lh_caudalanteriorcingulate_thickness	0.08	147	0.10	153
Right_Caudate	0.18	59	0.10	154
lh_supramarginal_volume	0.07	162	0.10	155
lh_precentral_volume	0.08	156	0.10	156
lh_temporalpole_thickness	0.04	169	0.10	157
lh_rostralmiddlefrontal_thickness	0.16	73	0.10	158
rh_rostralmiddlefrontal_thickness	0.15	82	0.10	159
rh_precuneus_thickness	0.19	53	0.10	160
lh_caudalmiddlefrontal_thickness	0.13	98	0.10	161
lh_parstriangularis_thickness	0.32	18	0.09	162
rh_parsorbitalis_volume	0.10	119	0.09	163
rh_pericalcarine_volume	0.13	96	0.09	164
lh_rostralmiddlefrontal_volume	0.08	149	0.09	165
lh_parsorbitalis_volume	0.17	68	0.09	166
lh_caudalmiddlefrontal_volume	0.08	153	0.09	167
Left_Caudate	0.12	100	0.08	168
lh_precentral_thickness	0.12	105	0.07	169

sTable 6.4. Coefficients and ranking for each feature for the pooled and cross-site validations for the SVM model (r_t=.73, p<.001).

	Cross-site validation		Pooled validation	
	Coefficients	Rank	Coefficients	Rank
Rh_inferiorparietal_thickness	0.86	1	0.85	1
Rh_insula_volume	0.81	3	0.83	2
Right_Lateral_Ventricle	0.63	6	0.72	3
Right_Hippocampus	0.82	2	0.72	4
Right_Inf_Lat_Vent	0.67	4	0.69	5
3rd_Ventricle	0.63	5	0.67	6
Lh_transversetemporal_thickness	0.54	10	0.66	7
Lh_insula_volume	0.53	11	0.66	8
Right_Putamen	0.62	7	0.61	9
Lh_parstriangularis_volume	0.52	12	0.60	10
Lh_parstriangularis_thickness	0.52	14	0.57	11
Left_Cerebellum_White_Matter	0.61	8	0.56	12
Rh_precuneus_thickness	0.44	24	0.54	13
Lh_rostralanteriorcingulate_thickness	0.58	9	0.51	14
Lh_precuneus_thickness	0.44	23	0.50	15
Lh_medialorbitofrontal_thickness	0.38	33	0.50	16
Lh_superiorfrontal_thickness	0.49	18	0.50	17
Rh_superiorfrontal_volume	0.46	21	0.50	18
Rh_middletemporal_volume	0.51	17	0.47	19
Rh_supramarginal_thickness	0.47	19	0.46	20
Lh_parahippocampal_thickness	0.40	29	0.45	21
Left_Pallidum	0.38	34	0.45	22
Lh_middletemporal_thickness	0.46	22	0.45	23
Rh_rostralanteriorcingulate_volume	0.40	28	0.45	24
Lh_temporalpole_volume	0.51	15	0.43	25
CC_Central	0.52	13	0.43	26
Rh_caudalanteriorcingulate_volume	0.33	44	0.43	27
Rh_lingual_thickness	0.41	25	0.43	28
Lh_lateralorbitofrontal_thickness	0.51	16	0.43	29
Lh_inferiorparietal_thickness	0.32	49	0.42	30
Rh_medialorbitofrontal_thickness	0.41	27	0.42	31
Lh_parsopercularis_volume	0.41	26	0.42	32
Rh_lateralorbitofrontal_thickness	0.33	46	0.38	33
Lh_inferiortemporal_volume	0.28	58	0.38	34
Lh_inferiorparietal_volume	0.39	32	0.37	35
Lh_transversetemporal_volume	0.25	65	0.37	36
Left_Thalamus_Proper	0.46	20	0.37	37
Rh_middletemporal_thickness	0.37	35	0.36	38
Lh_lingual_thickness	0.34	40	0.35	39
Rh_transversetemporal_volume	0.28	57	0.35	40
Rh_postcentral_thickness	0.34	41	0.34	41
Rh_parahippocampal_thickness	0.29	52	0.34	42
Lh_medialorbitofrontal_volume	0.24	67	0.34	43
Left_Lateral_Ventricle	0.39	31	0.33	44
Lh_inferiortemporal_thickness	0.34	42	0.31	45
Right_Accumbens_area	0.35	38	0.31	46
Right_Amygdala	0.36	36	0.30	47
Lh_middletemporal_volume	0.30	51	0.30	48
Right_Cerebellum_White_Matter	0.19	81	0.29	49
Lh_paracentral_thickness	0.29	53	0.29	50
Rh_posteriorcingulate_volume	0.35	39	0.29	51
Lh_precuneus_volume	0.27	61	0.28	52
Rh_entorhinal_thickness	0.34	43	0.28	53
Lh_caudalmiddlefrontal_volume	0.26	62	0.28	54
Lh_lateraloccipital_thickness	0.11	103	0.27	55
Rh_pericalcarine_thickness	0.39	30	0.26	56
CC_Mid_Posterior	0.19	76	0.25	57
Rh_precentral_thickness	0.32	48	0.24	58
Rh_isthmuscingulate_volume	0.29	56	0.24	59
Lh_superiortemporal_thickness	0.33	45	0.23	60
Right_ventraldc	0.25	64	0.23	61
Lh_cuneus_thickness	0.26	63	0.23	62
Rh_caudalmiddlefrontal_volume	0.23	69	0.22	63

Rh_parstriangularis_thickness	0.18	84	0.22	64
Rh_lateraloccipital_volume	0.27	60	0.21	65
Rh_cuneus_volume	0.21	72	0.21	66
Rh_transversetemporal_thickness	0.19	80	0.21	67
Rh_postcentral_volume	0.29	54	0.21	68
Lh_temporalpole_thickness	0.19	78	0.21	69
Rh_frontalpole_volume	0.21	73	0.20	70
Left_Inf_Lat_Vent	0.27	59	0.20	71
Rh_rostralanteriorcingulate_thickness	0.16	90	0.19	72
Rh_superiorparietal_thickness	0.17	86	0.19	73
Rh_parstriangularis_volume	0.11	105	0.19	74
Lh_postcentral_volume	0.19	79	0.19	75
Lh_entorhinal_thickness	0.25	66	0.19	76
Rh_lateraloccipital_thickness	0.30	50	0.19	77
Lh_parahippocampal_volume	0.16	92	0.18	78
Right_Thalamus_Proper	0.22	71	0.18	79
Rh_fusiform_thickness	0.19	77	0.18	80
Rh_superiortemporal_volume	0.24	68	0.18	81
Rh_temporalpole_thickness	0.36	37	0.17	82
Lh_caudalanteriorcingulate_thickness	0.15	93	0.17	83
Lh_posteriorcingulate_thickness	0.18	83	0.16	84
Left_Putamen	0.14	97	0.16	85
Lh_entorhinal_volume	0.32	47	0.16	86
Left_Cerebellum_Cortex	0.20	75	0.16	87
Rh_caudalanteriorcingulate_thickness	0.07	127	0.16	88
Rh_bankssts_volume	0.20	74	0.15	89
Lh_insula_thickness	0.29	55	0.15	90
Lh_fusiform_thickness	0.02	159	0.15	91
Lh_parsorbitalis_volume	0.07	124	0.15	92
Rh_lateralorbitofrontal_volume	0.22	70	0.14	93
Lh_cuneus_volume	0.13	99	0.14	94
Rh_lingual_volume	0.07	126	0.14	95
Rh_inferiortemporal_thickness	0.08	123	0.14	96
Rh_parsorbitalis_thickness	0.02	158	0.14	97
Rh_inferiortemporal_volume	0.03	156	0.14	98
Lh_rostralmiddlefrontal_thickness	0.15	96	0.14	99
Lh_precentral_volume	0.17	85	0.13	100
Rh_inferiorparietal_volume	0.03	152	0.13	101
Brain_Stem	0.11	104	0.13	102
Rh_caudalmiddlefrontal_thickness	0.16	91	0.13	103
Rh_superiorparietal_volume	0.08	119	0.13	104
Left_Hippocampus	0.11	108	0.13	105
Lh_rostralmiddlefrontal_volume	0.09	118	0.13	106
Rh_precentral_volume	0.06	135	0.13	107
Lh_lateraloccipital_volume	0.10	110	0.12	108
CC_Mid_Anterior	0.09	113	0.12	109
Lh_pericalcarine_volume	0.12	100	0.12	110
Rh_bankssts_thickness	0.16	88	0.12	111
Lh_bankssts_volume	0.06	133	0.12	112
Lh_parsorbitalis_thickness	0.16	89	0.12	113
Lh_postcentral_thickness	0.05	142	0.12	114
Rh_medialorbitofrontal_volume	0.11	102	0.12	115
Lh_frontalpole_thickness	0.09	116	0.11	116
Rh_frontalpole_thickness	0.16	87	0.11	117
Rh_posteriorcingulate_thickness	0.14	98	0.11	118
Left_ventraldc	0.08	122	0.10	119
Left_Accumbens_area	0.04	150	0.10	120
Rh_rostralmiddlefrontal_thickness	0.02	161	0.09	121
Lh_frontalpole_volume	0.05	136	0.09	122
Lh_superiorparietal_thickness	0.02	160	0.09	123
Lh_precentral_thickness	0.01	166	0.09	124
Rh_supramarginal_volume	0.19	82	0.09	125
Rh_paracentral_thickness	0.04	148	0.09	126
Lh_pericalcarine_thickness	0.09	115	0.09	127
Rh_parsopercularis_thickness	0.05	137	0.08	128
Rh_fusiform_volume	0.02	157	0.08	129
Rh_pericalcarine_volume	0.08	120	0.08	130
Lh_caudalanteriorcingulate_volume	0.07	125	0.07	131
Right_Caudate	0.05	141	0.07	132

CSF	0.04	147	0.07	133
Rh_isthmuscingulate_thickness	0.10	109	0.07	134
Lh_supramarginal_volume	0.08	121	0.07	135
Rh_superiortemporal_thickness	0.11	107	0.06	136
Lh_paracentral_volume	0.15	95	0.06	137
Lh_isthmuscingulate_volume	0.11	106	0.06	138
Left_Amygdala	0.09	112	0.06	139
Lh_supramarginal_thickness	0.00	168	0.06	140
Rh_insula_thickness	0.04	151	0.05	141
Lh_bankssts_thickness	0.05	138	0.05	142
Lh_rostralanteriorcingulate_volume	0.04	149	0.04	143
Lh_lingual_volume	0.01	164	0.04	144
Lh_fusiform_volume	0.09	114	0.04	145
Lh_superiortemporal_volume	0.04	144	0.04	146
Lh_lateralorbitofrontal_volume	0.06	129	0.04	147
Lh_superiorparietal_volume	0.05	139	0.03	148
Lh_parsopercularis_thickness	0.09	117	0.03	149
Rh_rostralmiddlefrontal_volume	0.06	132	0.03	150
Rh_parsorbitalis_volume	0.01	163	0.03	151
Lh_isthmuscingulate_thickness	0.04	145	0.03	152
Rh_precuneus_volume	0.05	143	0.02	153
CC_Posterior	0.05	140	0.02	154
Right_Pallidum	0.01	167	0.02	155
Rh_temporalpole_volume	0.09	111	0.02	156
Th_Ventricle	0.01	165	0.01	157
Rh_parsopercularis_volume	0.02	162	0.01	158
Rh_parahippocampal_volume	0.04	146	0.01	159
Left_Caudate	0.03	155	0.01	160
Rh_entorhinal_volume	0.00	169	0.01	161
Lh_posteriorcingulate_volume	0.03	153	0.01	162
Rh_paracentral_volume	0.06	134	0.01	163
Rh_superiorfrontal_thickness	0.06	131	0.00	164
Lh_caudalmiddlefrontal_thickness	0.12	101	0.00	165
Rh_cuneus_thickness	0.03	154	0.00	166
Right_Cerebellum_Cortex	0.07	128	0.00	167
Lh_superiorfrontal_volume	0.15	94	0.00	168
CC_Anterior	0.06	130	0.00	169

sTable 6.5. Coefficients and ranking for each feature for the pooled and cross-site validations for the regularized logistic regression model (r_t=.61, p<.001).

	Cross-site validation		Pooled validation	
	Coefficients	Rank	Coefficients	Rank
Rh_inferiorparietal_thickness	0.86	1	0.90	1
Rh_insula_volume	0.82	2	0.89	2
Right_Inf_Lat_Vent	0.76	5	0.81	3
Lh_transversetemporal_thickness	0.54	19	0.76	4
3rd_Ventricle	0.62	12	0.75	5
Right_Lateral_Ventricle	0.81	3	0.75	6
Lh_insula_volume	0.53	21	0.73	7
Right_Putamen	0.64	11	0.71	8
Right_Hippocampus	0.80	4	0.68	9
Lh_superiorfrontal_thickness	0.72	6	0.63	10
Lh_lateralorbitofrontal_thickness	0.49	25	0.53	11
Lh_parstriangularis_thickness	0.66	9	0.52	12
Right_Amygdala	0.35	51	0.51	13
Lh_precuneus_thickness	0.45	31	0.51	14
Rh_precuneus_thickness	0.46	30	0.51	15
Lh_rostralanteriorcingulate_thickness	0.70	7	0.50	16
Lh_medialorbitofrontal_thickness	0.58	18	0.49	17
Lh_middletemporal_thickness	0.52	23	0.49	18
Lh_parstriangularis_volume	0.68	8	0.48	19
Rh_parahippocampal_thickness	0.39	40	0.48	20
Rh_caudalanteriorcingulate_volume	0.61	14	0.48	21
Rh_rostralanteriorcingulate_volume	0.62	13	0.47	22
Rh_superiorfrontal_volume	0.59	15	0.47	23
Rh_middletemporal_volume	0.59	16	0.46	24
Lh_lingual_thickness	0.27	65	0.46	25
Lh_temporalpole_volume	0.50	24	0.45	26
CC_Central	0.42	34	0.45	27
Left_Cerebellum_White_Matter	0.64	10	0.45	28
Lh_inferiortemporal_volume	0.40	39	0.45	29
Lh_inferiorparietal_thickness	0.41	37	0.44	30
Lh_caudalmiddlefrontal_volume	0.31	61	0.43	31
Rh_supramarginal_thickness	0.54	20	0.42	32
Rh_lingual_thickness	0.38	45	0.39	33
Right_Cerebellum_White_Matter	0.17	93	0.38	34
Left_Thalamus_Proper	0.59	17	0.37	35
Rh_medialorbitofrontal_thickness	0.41	36	0.37	36
Left_Pallidum	0.31	60	0.37	37
Lh_parahippocampal_thickness	0.22	79	0.37	38
Lh_inferiorparietal_volume	0.39	41	0.37	39
Left_Hippocampus	0.32	58	0.36	40
Rh_postcentral_thickness	0.38	44	0.35	41
Lh_middletemporal_volume	0.53	22	0.34	42
Rh_pericalcarine_thickness	0.36	47	0.34	43
Right_Accumbens_area	0.43	33	0.34	44
Rh_isthmuscingulate_volume	0.36	48	0.32	45
Lh_medialorbitofrontal_volume	0.34	56	0.32	46
Lh_parsopercularis_volume	0.41	38	0.32	47
Lh_cuneus_thickness	0.06	137	0.32	48
Rh_lateralorbitofrontal_thickness	0.26	67	0.31	49
Lh_insula_thickness	0.35	52	0.31	50
Left_Lateral_Ventricle	0.47	29	0.31	51
Rh_transversetemporal_volume	0.44	32	0.30	52
Lh_inferiortemporal_thickness	0.34	54	0.30	53
Lh_posteriorcingulate_thickness	0.32	57	0.29	54
Lh_transversetemporal_volume	0.34	55	0.28	55
Rh_precentral_thickness	0.37	46	0.28	56
Lh_precuneus_volume	0.22	76	0.28	57
Rh_entorhinal_thickness	0.39	42	0.27	58
Rh_lateraloccipital_volume	0.32	59	0.27	59
Rh_bankssts_volume	0.14	104	0.26	60
Rh_superiorparietal_thickness	0.36	49	0.25	61
Rh_lateraloccipital_thickness	0.19	89	0.25	62
Lh_lateraloccipital_thickness	0.05	140	0.25	63

Lh_superiortemporal_thickness	0.47	27	0.25	64
Right_Thalamus_Proper	0.23	74	0.24	65
Lh_paracentral_thickness	0.36	50	0.24	66
Rh_rostralanteriorcingulate_thickness	0.35	53	0.24	67
Lh_temporalpole_thickness	0.22	77	0.24	68
Lh_rostralmiddlefrontal_volume	0.14	106	0.23	69
Lh_postcentral_volume	0.21	82	0.23	70
Rh_posteriorcingulate_volume	0.21	81	0.22	71
Lh_entorhinal_volume	0.17	94	0.22	72
Rh_postcentral_volume	0.38	43	0.21	73
Lh_caudalanteriorcingulate_thickness	0.12	110	0.20	74
Rh_posteriorcingulate_thickness	0.41	35	0.20	75
Rh_middletemporal_thickness	0.47	28	0.19	76
Lh_superiorparietal_thickness	0.09	121	0.19	77
Lh_fusiform_thickness	0.09	120	0.19	78
Lh_rostralmiddlefrontal_thickness	0.21	83	0.18	79
Rh_supramarginal_volume	0.24	70	0.18	80
Right_ventraldc	0.30	62	0.17	81
Lh_bankssts_volume	0.01	159	0.17	82
Rh_transversetemporal_thickness	0.19	90	0.17	83
Lh_cuneus_volume	0.24	71	0.17	84
Rh_caudalanteriorcingulate_thickness	0.07	134	0.16	85
Rh_precentral_volume	0.15	100	0.16	86
Lh_precentral_volume	0.23	75	0.16	87
Left_Accumbens_area	0.09	116	0.16	88
Lh_frontalpole_volume	0.09	117	0.16	89
Left_Inf_Lat_Vent	0.24	69	0.15	90
Rh_inferiortemporal_thickness	0.20	86	0.15	91
Rh_parsorbitalis_thickness	0.00	168	0.14	92
Rh_parstriangularis_thickness	0.22	78	0.14	93
Left_Putamen	0.09	115	0.14	94
Left_Cerebellum_Cortex	0.29	63	0.14	95
Lh_parsorbitalis_thickness	0.15	101	0.14	96
Lh_pericalcarine_thickness	0.23	72	0.14	97
Rh_caudalmiddlefrontal_volume	0.26	66	0.13	98
Rh_temporalpole_thickness	0.47	26	0.13	99
Rh_fusiform_thickness	0.18	92	0.13	100
Rh_lingual_volume	0.21	84	0.13	101
Rh_cuneus_volume	0.07	131	0.13	102
Rh_inferiortemporal_volume	0.09	118	0.13	103
Lh_lingual_volume	0.03	149	0.13	104
Rh_rostralmiddlefrontal_volume	0.17	95	0.12	105
Rh_superiortemporal_volume	0.28	64	0.12	106
Lh_lateraloccipital_volume	0.06	138	0.12	107
CSF	0.02	153	0.12	108
Lh_pericalcarine_volume	0.01	161	0.12	109
Rh_superiortemporal_thickness	0.02	156	0.12	110
Left_Amygdala	0.03	152	0.12	111
Rh_superiorparietal_volume	0.13	108	0.11	112
Rh_bankssts_thickness	0.01	162	0.11	113
Rh_caudalmiddlefrontal_thickness	0.00	165	0.11	114
Lh_parsorbitalis_volume	0.17	96	0.11	115
Rh_medialorbitofrontal_volume	0.23	73	0.11	116
Rh_frontalpole_thickness	0.06	139	0.10	117
Rh_precuneus_volume	0.03	147	0.10	118
Lh_precentral_thickness	0.15	103	0.10	119
Rh_inferiorparietal_volume	0.09	122	0.10	120
Rh_frontalpole_volume	0.13	109	0.10	121
Lh_frontalpole_thickness	0.07	132	0.10	122
Lh_posteriorcingulate_volume	0.12	111	0.09	123
Rh_pericalcarine_volume	0.02	158	0.09	124
Rh_rostralmiddlefrontal_thickness	0.02	155	0.09	125
Rh_paracentral_thickness	0.04	144	0.09	126
CC_Mid_Posterior	0.20	87	0.09	127
Lh_paracentral_volume	0.07	129	0.08	128
Lh_superiorparietal_volume	0.03	150	0.08	129
Rh_parsopercularis_volume	0.02	154	0.08	130
Lh_lateralorbitofrontal_volume	0.08	124	0.08	131
Lh_caudalanteriorcingulate_volume	0.15	102	0.07	132

Lh_entorhinal_thickness	0.17	97	0.07	133
Right_Caudate	0.11	114	0.07	134
Brain_Stem	0.08	126	0.07	135
Rh_temporalpole_volume	0.03	148	0.07	136
Rh_lateralorbitofrontal_volume	0.19	88	0.06	137
Lh_isthmuscingulate_volume	0.11	113	0.06	138
Lh_rostralanteriorcingulate_volume	0.04	146	0.06	139
Lh_supramarginal_thickness	0.06	136	0.06	140
CC_Mid_Anterior	0.25	68	0.05	141
Lh_fusiform_volume	0.07	127	0.05	142
Rh_entorhinal_volume	0.01	160	0.05	143
Rh_superiorfrontal_thickness	0.12	112	0.05	144
Rh_isthmuscingulate_thickness	0.15	98	0.04	145
Right_Cerebellum_Cortex	0.00	166	0.04	146
Lh_superiortemporal_volume	0.01	163	0.04	147
Lh_postcentral_thickness	0.22	80	0.04	148
Lh_bankssts_thickness	0.07	128	0.03	149
CC_Posterior	0.04	145	0.03	150
Rh_parstriangularis_volume	0.05	142	0.03	151
Lh_supramarginal_volume	0.00	169	0.03	152
Rh_parahippocampal_volume	0.00	167	0.02	153
Rh_paracentral_volume	0.05	141	0.02	154
CC_Anterior	0.07	130	0.02	155
Lh_parahippocampal_volume	0.18	91	0.02	156
Lh_parsopercularis_thickness	0.03	151	0.02	157
Rh_fusiform_volume	0.05	143	0.02	158
Th_Ventricle	0.13	107	0.02	159
Lh_caudalmiddlefrontal_thickness	0.14	105	0.02	160
Rh_cuneus_thickness	0.01	164	0.02	161
Right_Pallidum	0.08	125	0.02	162
Left_Caudate	0.09	123	0.01	163
Rh_parsorbitalis_volume	0.15	99	0.01	164
Rh_parsopercularis_thickness	0.07	135	0.01	165
Left_ventraldc	0.02	157	0.01	166
Lh_isthmuscingulate_thickness	0.07	133	0.00	167
Rh_insula_thickness	0.09	119	0.00	168
Lh_superiorfrontal_volume	0.20	85	0.00	169

Chapter 7

General discussion

7.1. Summary of main findings

Neuroanatomical abnormalities in schizophrenia have been well documented for the past four decades (Bora et al., 2011; Glahn et al., 2008). Although evidence suggests qualitatively similar, albeit less severe, changes in those who have experienced a recent FEP (Egerton et al., 2011), findings have been heterogenous (X. Gao et al., 2018; Shah et al., 2017). This may be partially explained by the increased risk of false positives (Button et al., 2013) and heterogeneous findings (IntHout et al., 2015) associated with small local studies that dominate the literature. This is in line with the ongoing concerns across the wider neuroscientific community regarding the failure of replication and reproducibility of findings (Anonymus, 2013) and subsequent calls for greater collaboration to build larger and more robust studies (Ferguson et al., 2014; Poldrack & Gorgolewski, 2014; Toga et al., 2015). In parallel with this movement, there is also a growing demand for clinically translatable research (Borgwardt & Fusar-Poli, 2012). Indeed, while neuroimaging has led to significant progress in the understanding of the neural correlates of psychosis, it has yet to make substantial impact in clinical practice (Dazzan, 2014; Prata et al., 2014; Woo et al., 2017). Machine learning promises to meet this demand by allowing inferences to be made at the level of the individual (Hastie et al., 2001). Initial attempts at using machine learning to distinguish FEP individuals from HC based on neuroanatomical data have yielded inconsistent findings (Kambeitz et al., 2015; Schnack & Kahn, 2016). This may be partially due to a combination of small sample sizes and less-than-rigorous methods (Nieuwenhuis et al., 2012; Schnack & Kahn, 2016; Varoquaux, 2017). In addition, while most evidence comes from well-established techniques such as SVM, a new approach known as deep learning is emerging as a promising development (LeCun et al., 2015). Contrary to traditional machine learning methods, deep learning is capable of finding highly abstract information which may be particularly useful for detecting the intricate and widespread pattern of neuroanatomical abnormalities in FEP.

Based on the above, the overarching aim of this doctoral thesis was to investigate neuroanatomical abnormalities in FEP at group and individual level in a mega-analytic study. This was achieved by collating data from five previous and independent studies carried out at four separate research sites. This resulted in the largest sample of FEP individuals to be analysed so far. In this context, FEP and HC groups were first compared using a standard mass-univariate

approach to test for neuroanatomical alterations common to the five independent sites. This allowed to move beyond local sample-dependent findings by testing for the presence of abnormalities that are consistent across different sites. Next, a series of studies were conducted to investigate the predictive power of machine learning, and deep learning in particular, in discriminating FEP from HC using neuroanatomical information. First, a thorough review of the literature was carried out to survey the current evidence for deep learning in psychiatric and neurologic neuroimaging. Next, a deep learning model, along with three other standard and well-established methods for comparison, were used to discriminate FEP and HC at the individual level. Critically, the same models were applied to the five independent datasets separately to test for the reproducibility of findings. Finally, deep learning was used to classify FEP and HC after combining all datasets, in a large-scale mega-analysis. The main results from each study are summarized below in the context of the hypotheses outlined in Chapter 1.

Study 1: Neuroanatomical abnormalities in first episode psychosis across independent samples: a multi-centre mega-analysis

Consistent with H1, the results obtained through standard univariate analysis revealed a pattern of widespread GM volume reductions in fronto-temporal, insular and occipital regions bilaterally in FEP relative to HC. Some of the reductions, namely in the left gyrus rectus, medial orbital and inferior temporal gyri as well as in the right fusiform and lingual gyri, were negatively correlated with positive and negative symptoms, thus partially confirming H2. Similarly some of the reductions, namely in the right lingual gyrus and insula, were also negatively correlated with duration of illness, therefore partially confirming H3. Furthermore, these reductions were not associated with anti-psychotic medication, consistent with H4. An increase in GM volume in the right superior temporal gyrus was also found, although this was not associated with severity of psychotic symptoms, duration of illness or anti-psychotic medication.

Study 2: Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications

The review of deep learning applications to neuroimaging studies in psychiatric or neurological disorders published up until 1st August 2016 yielded a total of 25 studies. As with traditional

machine learning methods, the majority of applications have been diagnostic classification studies, followed by conversion to illness and prediction of treatment response. Most studies have been conducted in patients with MCI and AD mostly via the ADNI dataset. However, other disorders have also been analysed, namely ADHD, psychosis, epilepsy and cerebellar ataxia. Possibly due to the flexibility inherent to deep learning models and the novelty of this approach, the methodology of the studies varied considerably in terms of features used, preprocessing, feature engineering and deep learning architectures. With respect to the deep learning model used, most studies used some form of autoencoders, CNN or a combination of the two, or even DNNs. In the majority of studies comparing their deep learning model to another more well-established approach, typically a kernel-based method such as SVM, deep learning showed improved performance. However, given the novelty and popularity of deep learning, it is not possible to exclude the possibility of publication bias in favour of this method.

Study 3: Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence

The results from the series of DNN and traditional machine learning models applied to three different neuroanatomical features – VWGMV, VWCT and SB-ROIs – to classify FEP and HC at each of the five datasets proved partially unexcepted in the context of the results from previous studies (Kambeitz et al., 2015; Xiao et al., 2017). Specifically, accuracies ranged from 50% to 70% for SB ROIs; from 50% to 63% for VWGMV; and from 51% to 68% for VWCT, thus not confirming H5 and H6. The best accuracies (70%) were achieved in 2 of the 5 sites, when a DNN was applied to the SB-ROIs, therefore partially confirming H7 and H8. However, these models generalized poorly when tested on the remaining sites, with accuracies ranging between 50 and 55%. Upon attempting to interpreting these results, the pool of similar studies was tested for publication bias. Results showed a significant bias towards studies with inflated accuracies.

Study 4: Using deep learning and structural data to identify first-episode psychosis: a multi-centre mega-analysis

In the final study of this doctoral work, pooling the five independent datasets into one individual-level mega-analysis revealed that a DNN was able to discriminate between FEP and HC with an

accuracy of 65.4%, sensitivity of 63.3% and specificity of 67.5%, therefore not confirming H9. However, as hypothesised in H10, the DNN model outperformed, albeit by a small margin, the traditional machine learning approaches. When the model was trained in four sites and tested on the remaining site iteratively, accuracy was still statistically significant, albeit with a drop to 59.3%. Finally, as hypothesized in H10, there was a moderate level of agreement with respect to the brain regions contributing to classification for each model, suggesting a reproducible signature of neuroanatomical changes comprising the third and inferior lateral ventricles, basal ganglia (putamen and pallidum), temporal regions such as the hippocampus, entorhinal cortex and lingual gyrus, as well as the inferior parietal gyrus, orbitofrontal regions and the insula.

7.2. Relationship to previous work

Based on the above, four overarching themes emerged from this doctoral work: 1) neuroanatomical signature of FEP, 2) reliability and reproducibility of machine learning findings in psychosis, 3) the promise of deep learning 4) real-world application of machine learning models.

7.2.1. Neuroanatomical signature of first-episode psychosis

The large-scale group-level univariate analysis reported in Chapter 3 revealed a widespread pattern of GM reduction in orbitofrontal regions, several regions across the temporal cortex including in the superior, inferior and middle gyri, as well as in the lingual and fusiform gyri and the insula. This pattern of GM abnormalities differed, to some extent, to the one identified by the large-scale deep learning model in Chapter 6. Here, the main regions driving classification were the third and inferior lateral ventricles, basal ganglia including the putamen and pallidum, temporal regions such as the hippocampus, entorhinal cortex and lingual gyrus, as well as the inferior parietal gyrus, orbitofrontal regions and the insula. At least three reasons may explain this difference. First, as described in Chapter 2, VBM and SBM analysis rely on fundamentally different methods to extract different anatomical measures. While the former yields a metric (e.g. volume, thickness) at the level of the voxel, the latter was used in the present work to extract information at region-level according to a built-in atlas. Second, the VBM group-level analysis focused only on differences in GM volume, whereas the deep learning analysis used both volume and thickness as input features. Finally, and perhaps most importantly, group-level statistics and

machine learning address distinct questions. While in the former the goal is to find differences between groups as identified by a *p*-value, in the latter the aim is to classify each subject into groups by separating the two groups. From here it follows that, for the same sample and feature distribution, a highly significant group difference does not necessarily translate into a high classification accuracy, and vice-versa (Arbabshirani et al., 2017). However, it is interesting to note that some brain areas, namely orbitofrontal and temporal regions as well as the insula, emerged in both traditional VBM group- and machine learning analysis. In principle, a simultaneous significant group-level difference and high classification accuracy can arise when the mean value for a given feature is so far apart between the groups that the values of most of participants of the two groups do not overlap (Arbabshirani et al., 2017). This is an unlikely scenario since the overall accuracy, i.e. using all features, was modest. Therefore, it is more likely that the distribution of fronto- temporal-insular regions, regardless of how they were measured (i.e. with a whole-brain voxel or SM-ROI approach) for each group was just different enough to yield a significant difference in the VBM analysis and a larger weight compared to other regions in the machine learning analysis. The overlap in regions between the two methods may be partially explained by the fact that the group-level analysis was forced to find abnormalities in FEP common to several sites, in a more conservative analysis compared to simply analysing all sites together, thus allowing to mitigate site-specific differences. The fact that these regions emerged from such different analytical approaches and anatomical measurements suggests stability and reproducibility of these findings, possibly due to the use of a large sample, and provides further evidence for fronto-temporal-insular changes in the early stages of psychosis (X. Gao et al., 2018; Shah et al., 2017).

7.2.2. Reliability and reproducibility in machine learning in psychosis: sample size and heterogeneity

The poor to modest accuracies from the single-site analysis in Chapter 5 were initially unexpected. However, as discussed in that Chapter, approximately 60% of reported accuracies in FEP studies used a sample of 50 participants or less, which, in combination with the instability of such small samples (Nieuwenhuis et al., 2012), less-than-rigorous methods (Arbabshirani et al., 2017; Woo et al., 2017), higher risk of overfitting (Combrisson & Jerbi, 2015) and publication

bias shown in Chapter 5, may have resulted in a previously postulated (Schnack & Kahn, 2016; Woo et al., 2017) over-representation of inflated accuracies. The sample sizes for each site included in the present work can be considered large in comparison with most other studies of FEP on even ChSz (Kambeitz et al., 2015). However, a large variation of within-site performances (across algorithms and feature types) was still observed. This is somewhat in line with simulations of the effect of sample size on the stability of machine learning models in ChSz (Nieuwenhuis et al., 2012), albeit accuracies tend to be lower for FEP than established psychotic disorders. An additional symptom of instability in performance were the relatively large standard-deviations for each dataset (displayed in sFigure 5.3 in the supplementary materials in Chapter 5). In contrast, in the large-scale machine learning analyses combining all five datasets, the standard-deviations were, as expected, considerably smaller, indicating a more stable and reliable performance (Varoquaux, 2017). However, accuracies were also modest at best. This may seem counterintuitive at first, as in traditional statistics larger samples tend to equate to 'better' results. This is because the final result is defined by a *p*-value, which in turn depends on the sample size, such that even a small effect size can become significant in a large enough sample (Schnack, 2017). For example, in the mega mass-univariate analysis in Chapter 3, although the analysis was somewhat constrained by being forced to find differences between the groups common to all sites, such large sample could render small effects statistically significant. This raises important challenges regarding the use of the *p*-value in the era of Big Data in neuroimaging (Smith & Nichols, 2018). In supervised machine learning however, results rely on how separable groups are based on the degree of overlap between their respective feature distributions, which does not depend on sample size (Schnack, 2017). Therefore, as sample size increases, the main obstacle becomes finding alterations shared by a large number of individuals in one group in relation to the other group with an effect size large enough such that groups can be separated. From here it follows that, for single-site studies, sample size can be achieved at the expense of loosened inclusion criteria, which will result in a less homogenous sample. Here, finding a common pattern of alterations with a sufficient effect size will be more challenging. This is due to an increase in 'apparent' heterogeneity (Schnack, 2017), whereby the degree of heterogeneity is partially determined by how strict are the inclusion criteria. This may explain why the average results from site 2 (51.9%) and site 1 (61.8%) were the worst and best, respectively. While the former included

patients with any psychotic disorder, the former only recruited first-episode schizophrenia patients. Although such variation could have also been due to sampling, i.e. chance, it is possible that the distinct levels of 'apparent' heterogeneity may have played a role. Interestingly, the overall results on the remaining three sites were between sites 1 and 2, as were their recruitment criteria by including patients with non-affective psychosis. We can therefore speculate that lower performances were associated with loosened inclusion criteria, i.e. more heterogeneous samples, as anticipated by Schnack (2017).

Given the limited number of patients that can be recruited at a single site, the time-constraints imposed by funding bodies and the limited resources available, large-scale studies such as the one here can only be achieved, for now, by combining data from multiple large studies. This adds further heterogeneity, as different studies have different inclusion criteria and use different assessment protocols. Therefore, the heterogeneity that resulted from pooling all five studies is likely to have contributed to the modest performance in the mega machine learning analysis in Chapter 6. Nevertheless, this trade-off is thought to be beneficial in the long-run, since such a heterogeneous sample ensures that models do not learn to exploit site-specific nuisance variables predictive of, but not relevant to, the distinction between FEP and controls, ultimately leading to more generalizable models (Durstewitz et al., 2019). As more Big Data initiatives take place, the standardisation of these factors across studies may help mitigate this source of heterogeneity which could, in turn, maximise the modelling of the left-over 'true' heterogeneity of psychosis (Schnack, 2017), that is, the different clinical manifestations and neurobiological substrates across individuals (Brugger & Howes, 2017; Tordesillas-Gutierrez et al., 2015; Wolfers et al., 2018).

7.2.3. The promise of deep learning

As reported in Chapter 5, the best classification accuracy from the single-site analyses were achieved in two sites with DNNs. Specifically, these two DNN models were able to classify FEP and HC with 70% accuracy. This result is in line with several other similar studies (Kambeitz et al., 2015). However, both models generalized poorly when tested on the remaining four sites, a strong indicator of overfitting. While a good performance in independent sites is a common issue

in psychiatric imaging studies in general (Schnack & Kahn, 2016), generalization tends to be particularly difficult to achieve with deep learning models. Although a powerful approach that can, in principle, model any relationship, this comes at the expense of complexity. Therefore, as a n nonlinear approach with such a large number of parameters to estimate, deep learning is particularly prone to overfitting. As reported in Chapter 6, the DNN was also the best performing model when all sites were analysed together in a pooled validation, albeit with a modest accuracy overall. The (marginal) superior performance of DNN over traditional linear models suggests that the inter-relations between regions may be better captured with nonlinear associations. Nevertheless, the drop in performance when the same model was trained in all but one site, a more conservative estimate of generalizability, suggests that the initial performance may be partly due to overfitting. However, the performance of the deep learning model was still higher, although by a slim margin, than that of traditional linear classifiers. Overall, and in keeping with the conclusion from Chapter 4, deep learning performed at least the same or better than traditional machine learning models. Although the sample used in this study was much larger than previous studies of FEP, it may not have been large enough for such a complex approach as deep learning to capture the subtle neuroanatomical changes that characterize the early stages of psychosis without overfitting. Indeed, the birth of deep learning and the (still) growing popularity that quickly followed was propelled by the increasing availability of powerful computers and huge amounts of data that are still not available (and may never be) in psychiatric neuroimaging. For example, areas in which deep learning has excelled, such as image and speech analysis, typically use datasets with 10^6 examples (He, Zhang, Ren, & Sun, 2015; Krizhevsky et al., 2012). Therefore, the results of this doctoral work should be taken as an initial attempt to use this advanced method to identify abnormalities at the early stages of psychosis. As the Big Data movement keep gaining momentum, larger samples will allow to explore deep learning to its full potential and help asserting its superiority in psychiatric neuroimaging in general, and in psychosis in particular. The last decade has already witnessed an unprecedented increase in sample sizes psychosis research, from a few dozen to several hundred participants such as in this doctoral work. Studies with a few thousand participants have now also started to emerge in other major psychiatric disorders (Nunes et al., 2018; Wegmayr et al., 2018).

In addition to being especially ‘data hungry”, interpretability is an additional important caveat of deep learning networks (Hinton, 2018). Specifically, the retrieval of the features driving the decision-making of an algorithm is of particular importance to mental health and medicine in general. Here, a model that is capable of accurately diagnosing an individual or correctly predicting a relapse for example, without providing an explanation of how the decision was made, is likely to be received with caution or even suspicion. At the moment however, the strength of deep learning models lies mostly on its pure data-driven mechanistic predictions rather than an insightful view into the neurobiological mechanisms of psychiatric disorders (Durstewitz et al., 2019). This is because a network’s predictions are based on learned nonlinear features whose meaning depends on complex interactions between uninterpreted features from the previous layers (Hinton, 2018). The lack of transparency is a well-known property of these models and new methods to open the ‘black-box’ are emerging (Chakraborty et al., 2017). In this work, the brain regions driving classification in Chapter 6 have been previously implicated in psychosis, which provides some reassurance that the model made decision based on relevant information for the task. However, it is possible that, had the initial parameters used to initialize the network been different, the network could have derived other brain regions as important for classification. Similarly, once the model was trained and tested, had the method to extract the best contributing regions been different, the group of features identified could have also differed from the ones reported.

7.2.4. Real-world application of machine learning in early intervention services

Although much progress has been made since the initial studies in the early 2000s in ChSz patients, the deployment of machine learning based tools to clinical practice in early intervention services is yet to become reality. This section briefly discusses four issues and challenges that will likely come into focus in the years to come: i) model development and validation process, ii) how good is good enough?, iii) case-control design and reliability of diagnostic labels and iv) economic viability and ethical issues.

The model development and validation process

It has been suggested that for a model to be implemented in the real world, it will have to surpass

four increasingly demanding stages of validation, in which model performance is estimated in i) one sample using CV (development stage), ii) new independent datasets (prospective validation), iii) across multiple laboratories and scanners (generalization), iv) diverse populations. According to this framework, only 9% of neuroimaging-based models across the psychiatric spectrum had gone beyond the initial development phase up until 2016 (Woo et al., 2017). In ChSz, very few studies so far have attempted to validate their single-site models in independent samples (e.g. Nieuwenhuis et al., 2017). As expected, this figure is much lower in FEP, with only one study (Dluhoš et al., 2017), in addition to the present work, reaching the prospective validation stage. Critically, the performance in independent samples across the psychiatric spectrum has been remarkedly lower than studies in the model-development stage, suggesting substantial optimistic biases (Woo et al., 2017). Therefore, although machine learning has brought a new purpose to psychiatric neuroimaging, much work is still needed to achieve clinical translation. Much larger samples capable of capturing the 'true' heterogeneity of the early stages of psychosis, whilst putting in place strategies to mitigate sources of 'apparent' heterogeneity across research centres, combined with data- and model-sharing initiatives will be needed in order to progress beyond the development stage.

How good is good enough?

The implementation of a machine learning tool into the clinical practice will need to ensure a minimum required level of performance. This raises the question of how good is good enough. It has been suggested that a machine learning-based tool can be considered useful if it reaches an accuracy similar or superior to standard methods (Shortliffe & Sepúlveda, 2018). However, in the absence of robust biomarkers for psychosis (Prata et al., 2014), there is currently no benchmark from an existing biologically-based diagnostic method. Results from this doctoral work suggest that a reliable identification of FEP based on neuroanatomical data may be around 60%. Although it may be possible to improve this performance, accuracies in ChSz seem to be converging towards 70%, suggesting that the classification of FEP will probably not go beyond 70%. From here it follows that a machine learning-based tool to identify the initial stages of psychosis exclusively based on anatomical information may not be good enough. It is also important to consider the potential cost of misclassification. The selected threshold for a tool to identify FEP

should take into account the fact that the cost of erroneously misclassifying someone ill as healthy may be higher (e.g. further psychotic episodes, reliance on other carers, unemployment) than the cost of misclassifying someone healthy as ill (mainly unnecessary examinations). Thus, an algorithm which provides excellent sensitivity but only good specificity may be preferred to one with the opposite pattern (Savitz, Rauch, & Drevets, 2013).

The case-control design and reliability of diagnostic labels

The vast majority of machine learning studies in psychosis (Kambeitz et al., 2015; Zarogianni et al., 2013), including the present work, and in psychiatric neuroimaging in general (Arbabshirani et al., 2017; Wolfers et al., 2015) have been based on the case-control design, where a supervised algorithm attempts to separate participants diagnosed with a particular disorder from controls. Although intuitive, this application may only reach the same level of diagnostic accuracy as traditional, interview based, methods of clinical assessment. This is due to the fact that the initial development of the decision function and subsequent testing, relies on the distinction between subjects whose labels are pre-defined by the researcher. Therefore, it would be an expression of logical confusion to expect a supervised algorithm to allow better diagnostic classification than traditional clinical assessment from which it was developed. This circular logic is further compromised by the more fundamental limitation of the current diagnostic system. This system defines a series of diagnostic labels according to clusters of symptoms that can only be determined by clinical interviews and observation. This has resulted in diagnostic categories with low reliability, some even with an inter-rater agreement little better than chance (Freedman et al., 2013). The diagnosis of schizophrenia has shown an acceptable level of inter-rater agreement, which is far from optimal. In addition to low reliability, none of the current psychiatric diagnostic categories seem to have a corresponding neurobiological signature, which is indicative of poor biological validity (Kapur et al., 2012; Prata et al., 2014). Although these are well-recognized issues in psychiatry, its implications for machine learning applications have been less discussed. In diagnostic classification studies, such as this thesis, where the aim is to use some neurobiological measure to separate individuals with a predefined diagnostic label from disease-free individuals (defined by the absence of any diagnostic label), the success of classifiers is limited, in part, by the reliability and biological validity of the diagnostic labels. Specifically, we

assume the diagnostic labels behave as in Figure 7.1A. i.e. they measure what we want (they are valid) every time (they are reliable). However, based on the exposed above, some would argue that Figure 7.1B better illustrates the current scenario, i.e. psychiatric diagnostic labels may not be as reliable or valid.

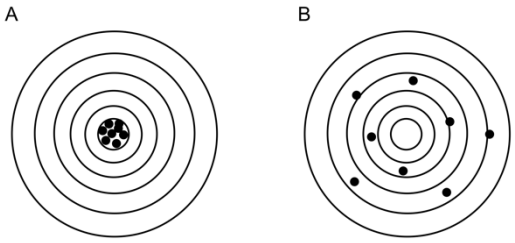


Figure 7.1. Bull's eye analogy to illustrate the validity and reliability of psychiatric diagnostic labels. The different dots represent different clinical assessments and respective diagnostic labels; the closer the dots are to each other, the higher the reliability. The centre represents what is meant to be measured, e.g. diagnosis of schizophrenia; the closer to the centre, the higher the validity. A represents a scenario where diagnostic labels are both reliable and valid. In B, diagnostic labels are unreliable and not valid.

Therefore, if, for example, some of the FEP individuals included in the present work have been wrongly labelled (and/or the controls wrongly labelled as not having any mental disorder), then finding a pattern that differentiates the two groups will become more difficult. Similarly, if FEP as we define it is not a valid neurobiological construct, then there may not be a robust pattern to be found. Taken collectively, these issues could also explain the modest results found in this thesis, as well as the results found in psychiatry in general. Moving forward, machine learning applications to psychiatry will have to circumvent these issues. Sophisticated tools, therefore, should be able to make decisions beyond the constraints of diagnostic labels ascertained by traditional assessments and criteria. This may be addressed in a number of ways including, using the extracted features maps and their specificities and commonalties across different psychiatric disorders to redefine current nosological systems, or alternatively omit labels all together and focus on the prediction of different functional domains as described in Research Domain Criteria (RDoC) (Cuthbert, 2015; Morris & Cuthbert, 2012) or use unsupervised methods to identify new demarcations between individuals (Durstewitz et al., 2019).

Economic viability and ethical issues

Before the implementation of any machine learning-based tool in clinical practice, a thorough assessment of cost-effectiveness and ethical implications will be necessary. At the moment, due to the absence of reliable biomarkers for functional psychosis (e.g. schizophrenia), MRI assessments in early intervention services are generally reserved for atypical cases and/or cases of suspected organic psychosis (e.g. epilepsy, tumour, encephalitis) (Borgwardt & Schmidt, 2017), a rare condition that can led to serious consequences if undetected (Joyce, 2018; Keshavan & Kaneko, 2013). Given its limited purpose, the routine use of neuroimaging has been shown to have limited economical cost-effectiveness (Albon et al., 2008; Khandanpour, Hoggard, & Connolly, 2013). However, the hope is that machine learning will greatly expand the use of imaging in early intervention services, not only for diagnostic but mostly for prognostic and treatment optimization purposes. Going forward, the economic burden of routine MRI assessments will have to be revisited to take into account the potential of machine learning in reducing the current elevated economic and societal burden associated with psychotic disorders (Olesen et al., 2012). Ethical issues will also have to be considered. Perhaps of special importance are the issues of accountability and privacy. As decision-making capacity is transferred from the human to the intelligent system, the moral and legal accountability of the human will diminish accordingly (Goering, Klein, Dougherty, & Widge, 2017; Kellmeyer et al., 2016). This will generate important discussions about how to create and implement legal guidelines to adjudicate accountability in cases of system failures. Similarly, the current trend of digitalization of biomedical data will result in enormous stockpiles of personal information, and dedicated efforts will have to be put in place to protect this data from unwarranted access and illegitimate use as well as to preserve the privacy of individual patients (Amunts, 2018; Kellmeyer, 2018; Yuste et al., 2017).

7.3. Strengths

Overall, the core strengths of the present doctoral work are fivefold. First, the sample size was the largest ever FEP sample of structural imaging data to be analysed with either univariate or multivariate methods. Indeed, the vast majority of neuroanatomical studies so far have been small local studies (X. Gao et al., 2018; Shah et al., 2017) which have been associated with a higher

risk of false positives (Button et al., 2013) and heterogeneous findings (IntHout et al., 2015). By combining five datasets, I was able to show, for the first time, a pattern of GM volume reductions related to symptoms severity that was present consistently across several independent sites in a group-level mega-analysis. This allowed to address site-dependent findings by identifying structural abnormalities above and beyond site-related differences. Similarly, the multivariate mega-analysis with traditional machine learning and deep learning provided a more reliable insight into the application of machine learning to neuroanatomical data in psychosis. Although accuracies were lower than that of most previous single-site studies (Kambeitz et al., 2015; Xiao et al., 2017), this was not unexpected given the increased heterogeneity inherent to larger and multi-centre studies (Schnack & Kahn, 2016). In addition, such a large sample is likely to be more representative of the FEP population and therefore should lead to results with more translational potential. Large samples sizes are also especially important for deep learning. Given their complexity and nonlinearity, these models thrive when applied to large amounts of data. While sample sizes similar to other areas of research are not yet possible, the work presented in this thesis was the largest deep learning investigation of the neuroanatomical basis of psychosis to date.

Second, as the application of deep learning keeps gains momentum across clinical neuroimaging, the first review of the current evidence in psychiatric and neurologic neuroimaging carried out as part of this doctoral work represents a useful resource for the neuroimaging community. The review also provides an introduction of deep learning to non-experts including its strengths but also its pitfalls such as proneness to overfitting and lack of transparency. Given the notorious popularity of deep learning, a thorough evaluation of the literature as well as its merits and limitations are essential to avoid misconceptions of its potential. This is especially relevant to non-experts, such as applied researchers and clinicians who are new to the field.

Third, I was able to show, through a series of carefully implemented multi -site, -feature and -machine learning algorithm experiments, that the predictive power of machine learning to recognise the initial stages of psychosis based on structural imaging may not be as high as initially thought. This comes at a time of growing concerns about potentially inflated findings due to the

use of small samples and less-than-rigours methods (Arbabshirani et al., 2017; Janssen et al., 2018; Woo et al., 2017). Indeed, most evidence so far has come from small local samples, where the possibility of a bespoke and likely overfitted pipeline cannot be ruled out. Consistent with this, I was also able to show evidence of publication bias, further supporting the already suspected over-representation of inflated results in the literature. Therefore, the work presented in this thesis is a valuable contribution towards recent calls for the next generation of machine learning studies with larger samples that allow more reliable estimates and independent sample validation (Schnack & Kahn, 2016).

Fourth, the use of deep neural networks represented the first attempt to classify FEP and HC based on neuroanatomical information using this novel approach. While, classic and simpler machine learning approaches are likely to be more suited for the traditional small neuroimaging study, as the Big Data movement gains momentum in the neuroimaging community, the combination of deep learning with large-scale samples may help capturing more complex interactions between neuroanatomical abnormalities across different brain regions.

Finally, by investigating individuals with a recent FEP, many of the confounding factors associated with ChSz are minimised if not removed completely, including the effects of prolonged exposure to anti-psychotic medication, effects of institutionalisation and the effects of chronicity. In addition, by focusing on those in the earliest stages of psychosis it is hoped that the findings may ultimately inform the provision of earlier and more effective treatment intervention, which in turn may delay the onset of psychotic relapse, if not prevent it altogether.

7.4. Limitations

A number of limitations also need to be considered when interpreting the results reported in this thesis. First, the structural MRI data was collected using different scanners and acquisition sequences. Although this was not an issue for the single-site analysis, it is most likely that this might have had an impact on the findings from the independent sample validations and mega-analyses in Chapters 5 and 6, respectively. For example, in the mega machine learning analysis, data from the different sites was harmonised by regressing out the effect of the scanner from each

feature. This is a standard method to deal with confounding variables in general (Rao, Monteiro, & Mourao-Miranda, 2017). However, perhaps a more sophisticated and tailored approach may have helped mitigating the effect of scanner even further.

Second, patients also differed with respect to exposure to anti-psychotic medication. Specifically, all patients were anti-psychotic naïve in site 1; patients from sites 3 and 4 were either naïve or had minimal exposure; finally, sites 2 and 5 had more relaxed criteria and included patients that met the inclusion criteria regardless of their anti-psychotic medication intake, resulting in a more heterogenous group. Although no effect was found in the univariate and single-site multivariate analyses, the impact of anti-psychotic medication cannot be ruled out. In addition, the possible effect of anti-psychotic medication was not tested for in the mega machine learning analysis in Chapter 6.

Third, the neuroanatomical abnormalities in FEP have been shown to vary according to several factors such as ethnicity (Gong et al., 2015), substance use (Rapp, Bugra, Riecher-Rossler, Tamagni, & Borgwardt, 2012) and IQ (Czepielewski, Wang, Gama, & Barch, 2017), for example. The studies included in the present work did not take these factors into account. This information was either not collected or made available for analyses. Had they been made available however, dealing with confounders in machine learning applied to neuroimaging is not trivial (Rao et al., 2017) and most studies deal with this issue by matching groups with respect to the key confounders variables.

7.5. Future work

Based on the findings presented here, there are a number of promising initiatives that could be developed as well as possible avenues to be investigated in years to come. Perhaps most vital is the standardization of several measurements that can facilitate the integration of the data from different sites. Although a data sharing movement is already underway in the psychiatric neuroimaging community, the benefits of having access to large amounts of data will have to be balanced against the issue of heterogeneity in assessment protocols, especially with respect to imaging acquisition. The uniformization of imaging protocols will allow pooling different datasets

into a single investigation with minimal impact of between-centre differences. On this regard, the ADNI consortium has developed a structural MRI sequence that generates an image with similar properties independently of the scanner model and manufacturer (Jack, Bernstein et al. 2008). While there are already a few multi-centre projects in psychosis that have adopted this approach (e.g. van Os et al. 2014; Cannon, Cadenhead et al. 2008), these are still a minority.

This doctoral work set out to discriminate FEP from HC using only neuroanatomical data. This imaging modality data was chosen based on the amount of data available which was considerably larger than that of other modalities. However, while maximizing sample size was a priority, there is increasing evidence suggesting that the neural mechanisms underlying psychosis may be better described as a complex pattern of 'dysconnectivity' in function and structure between brain regions (Friston, Brown, Siemerkus, & Stephan, 2016; Friston & Frith, 1995; McGuire & Frith, 1996; Pettersson-Yeo, Allen, Benetti, McGuire, & Mechelli, 2011). There is also evidence showing that this 'dysconnectivity hypothesis' is already present at the first psychotic episode (O'Neill et al., 2018). Although such evidence comes mostly from classic group-comparisons between patients and controls, a recent meta-analysis of machine learning studies in psychosis has also demonstrated the superiority of functional relative to neuroanatomical imaging data (Kambeitz et al., 2015), suggesting that brain connectivity may be more informative to identify psychosis at the individual level. At least one study has applied a DNN to functional connectivity to distinguish ChSz from HC, albeit in a relatively small sample of 100 participants (J. Kim et al., 2016) . Similar efforts dedicated to the initial stages of the illness in large-scale samples such as the one used in the present work could potentially shed light on the usefulness of functional connectivity without the typical confounds in ChSz. Additionally, one could also leverage on the different information conveyed by separate modalities in a multimodal approach (Calhoun & Sui, 2016). Deep learning may be particularly useful here, since we lack strong hypotheses of how different modalities interact (Durstewitz et al., 2019; Plis et al., 2018; Srinivasagopalan, Barry, Gurupur, & Thankachan, 2019).

The work presented here used the general-purpose and simplest form of deep leaning models. This was done due to limited computational resources. Future studies could make use of more

sophisticated networks such as autoencoders or convolutional networks for example, that may potentially unveil more useful patterns from single and/or multimodal imaging to identify FEP. Such approaches may be more computationally demanding and prone to overfitting. However, one may increase the effective “n” and alleviate computational burden simultaneously by using transfer learning, where knowledge gained with one dataset (or task) is transferring to another (Caruana, 1998). The main premise here is that distinct imaging datasets share basic structural properties and that leveraging on the hierarchical structure of the learned features with a deep learning model it is possible to use one dataset to learn more general properties of the data and then fine-tune the model in the dataset of interest. The increasing availability of imaging data will certainly be a useful resource in the next few years that may allow building models that capture the general essence of neuroimaging data, which are subsequently applied to specific tasks.

As with the vast majority of studies so far, the present work used the traditional case-control design to identify FEP individuals from HC. Despite intuitive, this design forces the algorithm to find a decision function that separates the two groups, which, given the known heterogeneity within both patients and controls (Brugger & Howes, 2017; Meyer-Lindenberg, 2010), may not be the most suitable approach to capture individualized patterns of abnormality needed for clinical translation (Marquand et al., 2016). Normative approaches, where individuals are mapped against a normative model that should encompass the heterogeneity characteristic of the normal population, are a promising avenue to address this issue (Marquand et al., 2016; Mourão-Miranda et al., 2011). Indeed, by considering illness as an extreme case within a normal range, the normative approach lends itself better to the already existing notion amongst clinicians that psychopathology is better understood as an extreme of a ‘normal’ phenomenon. The combined use of normative models with the increasing availability of data is a promising opportunity for further developments in machine learning-based tools for clinical decision making (Pinaya, Mechelli, & Sato, 2018).

Finally, future studies could also expand the approach used in this present work to include other diagnosis and/or investigate longitudinal outcomes. Distinguishing between alternative diagnoses, anticipating relapses or recovery, and choosing the optimal treatment for a particular

patient are likely to be of more clinical value than distinguishing between patients with FEP from disease-free individuals. As sample sizes increase and new methods, such as the normative approach, bring machine learning closer to clinical decision making, the investigation of such outcomes will certainly be at the centre of psychiatric neuroimaging research in the next years to come.

7.6. Conclusion

The present doctoral work aimed to address recent calls for more reliable, reproducible and translatable findings in psychiatric neuroimaging. Overall, results showed that FEP individuals manifest volumetric changes in fronto-temporal-insular regions that can be detected both at the group and individual level. Furthermore, this thesis provided a more realistic picture of the potential of machine learning to differentiate the initial stages of psychosis from HC. Based on the collection of results presented in this thesis, including both single and multi-study analysis, I speculate that the reliable separation of FEP from controls at the individual level based on neuroanatomical information is around 60%. This is lower than the accuracies reported by the majority of previous small local studies. Finally, deep learning showed a marginal superiority over traditional methods, indicating promise for unravelling biomarkers for the early stages of psychosis. However, there are still important challenges to overcome. These include the expensive computational resources and the amount of data required. However, solutions to both these obstacles are evolving rapidly and Big Data in combination with deep learning models will likely play an important role in the quest for transitional tools in the next years to come.

References

Aas, M., Dazzan, P., Mondelli, V., Melle, I., Murray, R. M., & Pariante, C. M. (2014). A Systematic Review of Cognitive Function in First-Episode Psychosis, Including a Discussion on Childhood Trauma, Stress, and Inflammation. *Frontiers in Psychiatry*, *4*, 182. https://doi.org/10.3389/fpsyt.2013.00182

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Retrieved from http://arxiv.org/abs/1603.04467

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). https://doi.org/10.1145/2976749.2978318

Adler, C. M., Levine, A. D., DelBello, M. P., & Strakowski, S. M. (2005). Changes in Gray Matter Volume in Patients with Bipolar Disorder. *Biological Psychiatry*, *58*(2), 151–157. https://doi.org/10.1016/J.BIOPSYCH.2005.03.022

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *ArXiv Preprint ArXiv:1610.01644*.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., … Kindermans, P.-J. (2018). iNNvestigate neural networks! Retrieved from http://arxiv.org/abs/1808.04260

Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V, & Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *Journal of General Internal Medicine*, *19*(5 Pt 1), 460–465. https://doi.org/10.1111/j.1525-1497.2004.30091.x

Albon, E., Tsourapas, A., Frew, E., Davenport, C., Oyebode, F., Bayliss, S., … Meads, C. (2008). Structural neuroimaging in psychosis: a systematic review and economic evaluation. In *NIHR Health Technology Assessment Programme: Executive summaries*. NIHR Journals Library.

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

American Psychiatric Association. (2013). DSM-5: Diagnostic and statistical manual of mental disorders . Washington, DC: Author.

Amunts, K. (2018). Big-data studies need to be part of policy discussion. *Nature Human Behaviour*, *2*(2), 94.

Andreasen, N. C., Flaum, M., & Arndt, S. (1992). The Comprehensive Assessment of Symptoms and History (CASH). *Archives of General Psychiatry*, *49*(8), 615. https://doi.org/10.1001/archpsyc.1992.01820080023004

Anonymus. (2013). Announcement: Reducing our irreproducibility. *Nature*, *496*, 398. https://doi.org/10.1038/496398a

APA. (2000). *Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV-TR)*. Washington, DC: American Psychiatric Association.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.079

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(0), 40–79. https://doi.org/10.1214/09-SS054

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2000). Voxel-Based Morphometry—The Methods. *NeuroImage*, *11*(6), 805–821. https://doi.org/10.1006/NIMG.2000.0582

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018

Asmal, L., du Plessis, S., Vink, M., Chiliza, B., Kilian, S., & Emsley, R. (2018). Symptom attribution and frontal cortical thickness in first-episode schizophrenia. *Early Intervention in Psychiatry*, *12*(4), 652–659. https://doi.org/10.1111/eip.12358

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics* (pp. 26–33).

Barkl, S. J., Lah, S., Harris, A. W. F., & Williams, L. M. (2014). Facial emotion identification in early-onset and first-episode psychosis: A systematic review with meta-analysis. *Schizophrenia Research*, *159*(1), 62–69. https://doi.org/10.1016/J.SCHRES.2014.07.049

Bearden, C. E., & Thompson, P. M. (2017). Emerging Global Initiatives in Neurogenetics: The Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA) Consortium. *Neuron*,

94(2), 232–236. https://doi.org/10.1016/J.NEURON.2017.03.033

Bebbington, P., & Nayani, T. (1995). The psychosis screening questionnaire. *International Journal of Methods in Psychiatric Research*, *5*, 11–19.

Benetti, S., Pettersson-Yeo, W., Allen, P., Catani, M., Williams, S., Barsaglini, A., … Mechelli, A. (2015). Auditory Verbal Hallucinations and Brain Dysconnectivity in the Perisylvian Language Network: A Multimodal Investigation. *Schizophrenia Bulletin*, *41*(1), 192–200. https://doi.org/10.1093/schbul/sbt172

Benetti, S., Pettersson-Yeo, W., Hutton, C., Catani, M., Williams, S. C., Allen, P., … Mechelli, A. (2013). Elucidating neuroanatomical alterations in the at risk mental state and first episode psychosis: A combined voxel-based morphometry and voxel-based cortical thickness study. *Schizophrenia Research*, *150*(2–3), 505–511. https://doi.org/10.1016/J.SCHRES.2013.08.030

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (pp. 437–478). Springer.

Bengio, Y. (2009). *Learning Deep Architectures for AI*. *Foundations and Trends® in Machine Learning* (Vol. 2). https://doi.org/10.1561/2200000006

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). *Deep Learning*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.672.7118&rep=rep1&type=pdf

Berger, G. E., Wood, S., & McGorry, P. D. (2003). Incipient neurovulnerability and neuroprotection in early psychosis. *Psychopharmacology Bulletin*, *37*(2), 79–101. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14566217

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546–2554).

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, *107*(10), 4734–4739. https://doi.org/10.1073/pnas.0911855107

Blankstein, U., Chen, J. Y. W., Mincic, A. M., McGrath, P. A., & Davis, K. D. (2009). The complex minds of teenagers: neuroanatomy of personality differs between sexes. *Neuropsychologia*, *47*(2), 599–603. https://doi.org/10.1016/j.neuropsychologia.2008.10.014

Bora, E., Fornito, A., Radua, J., Walterfang, M., Seal, M., Wood, S. J., … Pantelis, C. (2011).

Neuroanatomical abnormalities in schizophrenia: A multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophrenia Research*, *127*(1–3), 46–57. https://doi.org/10.1016/J.SCHRES.2010.12.020

Borgwardt, S., & Fusar-Poli, P. (2012). Third-generation neuroimaging in early schizophrenia: Translating research evidence into clinical utility. *British Journal of Psychiatry*, *200*(4), 270–272. https://doi.org/10.1192/bjp.bp.111.103234

Borgwardt, S., Koutsouleris, N., Aston, J., Studerus, E., Smieskova, R., Riecher-Rossler, A., & Meisenzahl, E. M. (2013). Distinguishing Prodromal From First-Episode Psychosis Using Neuroanatomical Single-Subject Pattern Recognition. *Schizophrenia Bulletin*, *39*(5), 1105–1114. https://doi.org/10.1093/schbul/sbs095

Borgwardt, S., & Schmidt, A. (2017). Implementing magnetic resonance imaging into clinical routine screening in patients with psychosis? *The British Journal of Psychiatry*, *211*(4), 192–193.

Boulesteix, A.-L., Lauer, S., & Eugster, M. J. A. (2013). A Plea for Neutral Comparison Studies in Computational Sciences. *PLOS ONE*, *8*(4), 1–11. https://doi.org/10.1371/journal.pone.0061562

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition* (pp. 3121–3124). IEEE. https://doi.org/10.1109/ICPR.2010.764

Brosch, T., & Tam, R. (2013). Manifold Learning of Brain MRIs by Deep Learning BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, & N. Navab (Eds.) (pp. 633–640). Berlin, Heidelberg: Springer Berlin Heidelberg.

Brugger, S. P., & Howes, O. D. (2017). Heterogeneity and Homogeneity of Regional Brain Structure in Schizophrenia. *JAMA Psychiatry*, *74*(11), 1104. https://doi.org/10.1001/jamapsychiatry.2017.2663

Buchanan, R. W., Francis, A., Arango, C., Miller, K., Lefkowitz, D. M., McMahon, R. P., … Pearlson, G. D. (2004). Morphometric Assessment of the Heteromodal Association Cortex in Schizophrenia. *American Journal of Psychiatry*, *161*(2), 322–331. https://doi.org/10.1176/appi.ajp.161.2.322

Butler, P. D., Silverstein, S. M., & Dakin, S. C. (2008). Visual Perception and Its Impairment in

Schizophrenia. *Biological Psychiatry*, *64*(1), 40–47. https://doi.org/10.1016/J.BIOPSYCH.2008.03.023

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Bzdok, D. (2017). Classical Statistics and Statistical Learning in Imaging Neuroscience. *Frontiers in Neuroscience*, *11*, 543. https://doi.org/10.3389/fnins.2017.00543

Bzdok, D., & Yeo, B. T. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, *155*, 549–564. https://doi.org/10.1016/J.NEUROIMAGE.2017.04.061

Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V. D., Dwyer, D. B., von Saldern, S., … Koutsouleris, N. (2016). Classifying Schizophrenia Using Multimodal Multivariate Pattern Recognition Analysis: Evaluating the Impact of Individual Clinical Profiles on the Neurodiagnostic Performance. *Schizophrenia Bulletin*, *42*(suppl 1), S110–S117. https://doi.org/10.1093/schbul/sbw053

Calhoun, V. D., & Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *1*(3), 230–244. https://doi.org/10.1016/j.bpsc.2015.12.005

Caruana, R. (1998). Multitask Learning. In S. Thrun & L. Pratt (Eds.), *Learning to Learn* (pp. 95–133). Springer, Boston, MA.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., … Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1–6). IEEE. https://doi.org/10.1109/UIC-ATC.2017.8397411

Chan, R. C. K., Di, X., McAlonan, G. M., & Gong, Q. -y. (2011). Brain Anatomical Abnormalities in High-Risk Individuals, First-Episode, and Chronic Schizophrenia: An Activation Likelihood Estimation Meta-analysis of Illness Progression. *Schizophrenia Bulletin*, *37*(1), 177–188. https://doi.org/10.1093/schbul/sbp073

Chollet, F., & others. (2015). Keras.

Chua, S. E., Cheung, C., Cheung, V., Tsang, J. T. K., Chen, E. Y. H., Wong, J. C. H., … McAlonan, G. M. (2007). Cerebral grey, white matter and csf in never-medicated, first-episode schizophrenia. *Schizophrenia Research*, *89*(1–3), 12–21. https://doi.org/10.1016/j.schres.2006.09.009

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, *250*, 126–136.

Consortium, S. W. G. of the P. G., Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595

Contreras, N. A., Tan, E. J., Lee, S. J., Castle, D. J., & Rossell, S. L. (2018). Using visual processing training to enhance standard cognitive remediation outcomes in schizophrenia: A pilot study. *Psychiatry Research*, *262*, 494–499. https://doi.org/10.1016/J.PSYCHRES.2017.09.031

Cuthbert, B. N. (2015). Research Domain Criteria: toward future psychiatric nosologies. *Dialogues in Clinical Neuroscience*, *17*(1), 89.

Czepielewski, L. S., Wang, L., Gama, C. S., & Barch, D. M. (2017). The relationship of intellectual functioning and cognitive performance to brain structure in schizophrenia. *Schizophrenia Bulletin*, *43*(2), 355–364.

Dahne, S., Bieszmann, F., Samek, W., Haufe, S., Goltz, D., Gundlach, C., … Muller, K.-R. (2015). Multivariate Machine Learning Methods for Fusing Multimodal Functional Neuroimaging Data. *Proceedings of the IEEE*, *103*(9), 1507–1530. https://doi.org/10.1109/JPROC.2015.2425807

Dale, A. M., Fischl, B., & Sereno, M. I. (1999a). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dale, A. M., Fischl, B., & Sereno, M. I. (1999b). *Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction*. *NeuroImage* (Vol. 9). Academic Press. https://doi.org/10.1006/NIMG.1998.0395

Davatzikos, C., Shen, D., Gur, R. C. R. E., Wu, X., Liu, D., Fan, Y., … Gur, R. C. R. E. (2005). Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal

abnormalities. *Archives of General Psychiatry*, *62*(11), 1218–1227. https://doi.org/10.1001/archpsyc.62.11.1218

Dazzan, P. (2014). Neuroimaging biomarkers to predict treatment response in schizophrenia: the end of 30 years of solitude? *Dialogues in Clinical Neuroscience*, *16*(4), 491–503. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25733954

de Moura, A. M., Pinaya, W. H. L., Gadelha, A., Zugman, A., Noto, C., Cordeiro, Q., … Sato, J. R. (2018). Investigating brain structural patterns in first episode psychosis and schizophrenia using MRI and a machine learning approach. *Psychiatry Research: Neuroimaging*, *275*, 14–20. https://doi.org/10.1016/j.pscychresns.2018.03.003

de Pierrefeu, A., Löfstedt, T., Laidi, C., Hadj-Selem, F., Bourgin, J., Hajek, T., … Duchesnay, E. (2018). Identifying a neuroanatomical signature of schizophrenia, reproducible across sites and stages, using machine learning with structured sparsity. *Acta Psychiatrica Scandinavica*. https://doi.org/10.1111/acps.12964

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, *58*(9), 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016

der Werf, M., Hanssen, M., Köhler, S., Verkaaik, M., Verhey, F. R., van Winkel, R., … others. (2014). Systematic review and collaborative recalculation of 133 693 incident cases of schizophrenia. *Psychological Medicine*, *44*(1), 9–16.

Deshpande, G., Wang, P., Rangaprakash, D., & Wilamowski, B. (2015). Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data. *Cybernetics, IEEE Transactions On*, *PP*(99), 1. https://doi.org/10.1109/TCYB.2014.2379621

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Di Forti, M., Morgan, C., Dazzan, P., Pariante, C., Mondelli, V., Marques, T. R., … Murray, R. M. (2009). High-potency cannabis and the risk of psychosis. *British Journal of Psychiatry*, *195*(06), 488–491. https://doi.org/10.1192/bjp.bp.109.064220

Dluhoš, P., Schwarz, D., Cahn, W., van Haren, N., Kahn, R., Španiel, F., … Schnack, H. (2017). Multi-center machine learning in imaging psychiatry: A meta-model approach. *NeuroImage*, *155*, 10–24. https://doi.org/10.1016/J.NEUROIMAGE.2017.03.027

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 677–691. https://doi.org/10.1109/tpami.2016.2599174

Donini, M., Monteiro, J. M., Pontil, M., Hahn, T., Fallgatter, A. J., Shawe-Taylor, J., & Mourão-Miranda, J. (2019). Combining heterogeneous data sources for neuroimaging based diagnosis: re-weighting and selecting what is important. *NeuroImage*, *195*, 215–231. https://doi.org/10.1016/J.NEUROIMAGE.2019.01.053

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, 1. https://doi.org/10.1038/s41380-019-0365-9

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*(8), 2530–2541. https://doi.org/10.1002/hbm.20688

Egerton, A., Borgwardt, S. J., Tognin, S., Howes, O. D., McGuire, P., & Allen, P. (2011, October). An overview of functional, structural and neurochemical imaging studies in individuals with a clinical high risk for psychosis. *Neuropsychiatry*. https://doi.org/10.2217/npy.11.51

Eickhoff, S., Nichols, T. E., Van Horn, J. D., & Turner, J. A. (2016). Sharing the wealth: Neuroimaging data repositories. *NeuroImage*, *124*(Pt B), 1065–1068. https://doi.org/10.1016/j.neuroimage.2015.10.079

Ellison-Wright, I., & Bullmore, E. (2010). Anatomy of bipolar disorder and schizophrenia: A meta-analysis. *Schizophrenia Research*, *117*(1), 1–12. https://doi.org/10.1016/J.SCHRES.2009.12.022

Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M., & Bullmore, E. (2008). The Anatomy of First-Episode and Chronic Schizophrenia: An Anatomical Likelihood Estimation Meta-Analysis. *American Journal of Psychiatry*, *165*(8), 1015–1023. https://doi.org/10.1176/appi.ajp.2008.07101562

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., … Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fei, B., & Liu, J. (2006). Binary tree of SVM: a new fast multiclass training and classification algorithm. *IEEE Transactions on Neural Networks*, *17*(3), 696–704. https://doi.org/10.1109/TNN.2006.872343

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: data-sharing in the “long tail” of neuroscience, *17*(11), 1442–1447. https://doi.org/10.1038/nn.3838

Fett, A.-K. J., Viechtbauer, W., Dominguez, M.-G., Penn, D. L., van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 573–588. https://doi.org/10.1016/J.NEUBIOREV.2010.07.001

First, M. B., & Gibbon, M. (2004). The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II).

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/J.NEUROIMAGE.2012.01.021

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*, S69–S84. https://doi.org/10.1016/j.neuroimage.2004.07.016

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). *Cortical Surface-Based Analysis II: Inflation, Flattening, and a Surface-Based Coordinate System*. Retrieved from

http://www.idealibrary.com

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Fornito, A., Yücel, M., Wood, S. J., Adamson, C., Velakoulis, D., Saling, M. M., … Pantelis, C. (2008). Surface-based morphometry of the anterior cingulate cortex in first episode schizophrenia. *Human Brain Mapping*, *29*(4), 478–489. https://doi.org/10.1002/hbm.20412

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678. https://doi.org/10.1073/pnas.0504136102

Frackowiak, R. S. J. (2004). *Human brain function*. Elsevier.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., … others. (2013). The initial field trials of DSM-5: new blooms and old thorns. Am Psychiatric Assoc.

Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia Research*, *176*(2–3), 83–94. https://doi.org/10.1016/J.SCHRES.2016.07.014

Friston, K., & Frith, C. D. (1995). Schizophrenia: a disconnection syndrome. *Clin Neurosci*, *3*(2), 89–97.

Friston, K., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Fusar-Poli, P., Borgwardt, S., Crescini, A., Deste, G., Kempton, M. J., Lawrie, S., … Sacchetti, E. (2011). Neuroanatomy of vulnerability to psychosis: A voxel-based meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1175–1185. https://doi.org/10.1016/j.neubiorev.2010.12.005

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., … Politi, P. (2009). Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience : JPN*, *34*(6), 418–432. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19949718

Fusar-Poli, P., Radua, J., McGuire, P., & Borgwardt, S. (2012). Neuroanatomical Maps of Psychosis Onset: Voxel-wise Meta-Analysis of Antipsychotic-Naive VBM Studies. *Schizophrenia Bulletin*, *38*(6), 1297–1307. https://doi.org/10.1093/schbul/sbr134

Fusar-Poli, P., Smieskova, R., Serafini, G., Politi, P., & Borgwardt, S. (2014). Neuroanatomical markers of genetic liability to psychosis and first episode psychosis: A voxelwise meta-analytical comparison. *The World Journal of Biological Psychiatry*, *15*(3), 219–228. https://doi.org/10.3109/15622975.2011.630408

Gao, X. W., & Hui, R. (2016). A deep learning based approach to classification of CT brain images. In *2016 SAI Computing Conference (SAI)* (pp. 28–31). https://doi.org/10.1109/SAI.2016.7555958

Gao, X., Zhang, W., Yao, L., Xiao, Y., Liu, L., Liu, J., … Lui, S. (2018). Association between structural and functional brain alterations in drug-free patients with schizophrenia: a multimodal meta-analysis. *J Psychiatry Neurosci*, *43*(2), 131–142. https://doi.org/10.1503/jpn.160219

Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian optimization with unknown constraints. *ArXiv Preprint ArXiv:1403.5607*.

Glahn, D. C., Laird, A. R., Ellison-Wright, I., Thelen, S. M., Robinson, J. L., Lancaster, J. L., … Fox, P. T. (2008). Meta-Analysis of Gray Matter Anomalies in Schizophrenia: Application of Anatomic Likelihood Estimation and Network Analysis. *Biological Psychiatry*, *64*(9), 774–781. https://doi.org/10.1016/J.BIOPSYCH.2008.03.031

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

Goering, S., Klein, E., Dougherty, D. D., & Widge, A. S. (2017). Staying in the loop: Relational agency and identity in next-generation DBS for psychiatry. *AJOB Neuroscience*, *8*(2), 59–70.

Gong, Q., Dazzan, P., Scarpazza, C., Kasai, K., Hu, X., Marques, T. R., … Mechelli, A. (2015). A Neuroanatomical Signature for Schizophrenia Across Different Ethnic Groups. *Schizophrenia Bulletin*, *41*(6), 1266–1275. https://doi.org/10.1093/schbul/sbv109

Gong, Q., Li, L., Du, M., Pettersson-Yeo, W., Crossley, N., Yang, X., … Mechelli, A. (2014). Quantitative prediction of individual psychopathology in trauma survivors using resting-state FMRI. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *39*(3), 681–687. https://doi.org/10.1038/npp.2013.251

Gong, Q., Scarpazza, C., Dai, J., He, M., Xu, X., Shi, Y., … Mechelli, A. (2018). A transdiagnostic

neuroanatomical signature of psychiatric illness. *Neuropsychopharmacology*, 1. https://doi.org/10.1038/s41386-018-0175-9

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N. A., Friston, K. J., & Frackowiak, R. S. J. (2001). A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. https://doi.org/10.1006/nimg.2001.0786

Good, P. (1994). *Permutation Tests : a Practical Guide to Resampling Methods for Testing Hypotheses*. Springer New York.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Retrieved from https://www.deeplearningbook.org/

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). IEEE. https://doi.org/10.1109/ICASSP.2013.6638947

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620–631. https://doi.org/10.1038/nrn4005

Grün, F., Rupprecht, C., Navab, N., & Tombari, F. (2016). A taxonomy and library for visualizing learned features in convolutional neural networks. *ArXiv Preprint ArXiv:1606.07757*.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402. https://doi.org/10.1001/jama.2016.17216

Guo, S., Palaniyappan, L., Liddle, P. F., & Feng, J. (2016). Dynamic cerebral reorganization in the pathophysiology of schizophrenia: a MRI-derived cortical thickness study. *Psychological Medicine*, *46*(10), 2201–2214. https://doi.org/10.1017/S0033291716000994

Gupta, A., Ayhan, M. S., & Maida, A. S. (2013). Natural Image Bases to Represent Neuroimaging Data. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, *28*(3), 977–984.

Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., … Turner, J. A. (2015). Patterns of Gray Matter Abnormalities in Schizophrenia Based on an International Mega-analysis. *Schizophrenia Bulletin*, *41*(5), 1133–1142. https://doi.org/10.1093/schbul/sbu177

Gutiérrez-Maldonado, J., Caqueo-Urízar, A., & Kavanagh, D. J. (2005). Burden of care and general health in families of patients with schizophrenia. *Social Psychiatry and Psychiatric*

Epidemiology, *40*(11), 899–904. https://doi.org/10.1007/s00127-005-0963-5

Hahn, B., Ross, T. J., & Stein, E. A. (2006). Neuroanatomical dissociation between bottom–up and top–down processes of visuospatial selective attention. *NeuroImage*, *32*(2), 842–853. https://doi.org/10.1016/j.neuroimage.2006.04.177

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M. P., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain Volumes in Schizophrenia: A Meta-Analysis in Over 18 000 Subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138. https://doi.org/10.1093/schbul/sbs118

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., … Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194. https://doi.org/10.1016/J.NEUROIMAGE.2006.02.051

Han, Xiaobing, Zhong, Y., He, L., & Yu, P. S. (2015). The Unsupervised Hierarchical Convolutional Sparse Auto-Encoder for Neuroimaging Data Classification. *BIH*, *9250*, 156–166. https://doi.org/10.1007/978-3-319-23344-4

Hao, A. J., He, B. L., & Yin, C. H. (2015). Discrimination of ADHD children based on Deep Bayesian Network. *IET Conference Proceedings*, 6 .-6 .(1). Retrieved from https://digital-library.theiet.org/content/conferences/10.1049/cp.2015.0764

Haring, L., Müürsepp, A., Mõttus, R., Ilves, P., Koch, K., Uppin, K., … Vasar, V. (2016). Cortical thickness and surface area correlates with cognitive dysfunction among first-episode psychosis patients. *Psychological Medicine*, *46*(10), 2145–2155. https://doi.org/10.1017/S0033291716000684

Hastie, T. (2009). The Elements of Statistical Learning. *The Mathematical Intelligencer*, *27*(2), 83–85. https://doi.org/10.1007/b94608

Haxby, J. V, Hoffman, E. A., & Gobbini, M. I. (2002). *Human Neural Systems for Face Recognition and Social Communication*. *Biol Psychiatry* (Vol. 51). Retrieved from https://ac.els-cdn.com/S0006322301013300/1-s2.0-S0006322301013300-main.pdf?_tid=832ee091-e784-497a-bbad-0d006bc6ec73&acdnat=1536165522_f68360f7c271aaef3efe32f5f4cb3e41

Haxby, J. V, Hoffman, E. A., Gobbini, M. I., Haxby, J. V, Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences* –, *4*(6), 223–233. Retrieved from https://ac.els-cdn.com/S1364661300014820/1-s2.0-

S1364661300014820-main.pdf?_tid=04f6452d-140a-4455-bc57-8ab4859716ef&acdnat=1536165516_2c759bb9cc13916596b7f3334d7f8f74

Hayes, J. F., Marston, L., Walters, K., King, M. B., & Osborn, D. P. J. (2017). Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000–2014. *British Journal of Psychiatry*, *211*(03), 175–181. https://doi.org/10.1192/bjp.bp.117.202606

Hayes, L., Hawthorne, G., Farhall, J., O'Hanlon, B., & Harvey, C. (2015). Quality of Life and Social Isolation Among Caregivers of Adults with Schizophrenia: Policy and Outcomes. *Community Mental Health Journal*, *51*(5), 591–597. https://doi.org/10.1007/s10597-015-9848-6

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., … Network, T. I. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, *542*(7641), 348–351. https://doi.org/10.1038/nature21369

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Retrieved from http://arxiv.org/abs/1502.01852

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, *17*, 16–23. https://doi.org/10.1016/J.NICL.2017.08.017

Hibar, D. P., Westlye, L. T., Doan, N. T., Jahanshad, N., Cheung, J. W., Ching, C. R. K., … Andreassen, O. A. (2018). Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Molecular Psychiatry*, *23*(4), 932–942. https://doi.org/10.1038/mp.2017.73

Hinton, G. (2018). Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*, *320*(11), 1101. https://doi.org/10.1001/jama.2018.11100

Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *The Lancet Psychiatry*, *4*(4), 295–301. https://doi.org/10.1016/S2215-0366(17)30078-0

Holmes, A. J., & Patrick, L. M. (2018). The Myth of Optimality in Clinical Neuroscience. *Trends in Cognitive Sciences*, *22*(3), 241–257. https://doi.org/10.1016/J.TICS.2017.12.006

Honea, R., Crow, T. J., Passingham, D., & Mackay, C. E. (2005). Regional Deficits in Brain Volume in Schizophrenia: A Meta-Analysis of Voxel-Based Morphometry Studies. *American Journal of Psychiatry*, *162*(12), 2233–2245. https://doi.org/10.1176/appi.ajp.162.12.2233

Hosseini-Asl, E., Keynto, R., & El-Baz, A. (2016). Alzheimer's Disease Diagnostics by Adaptation of 3D Convolutional Network, (502). https://doi.org/10.1109/ICIP.2016.7532332

Howes, O., McCutcheon, R., & Stone, J. (2015). Glutamate and dopamine in schizophrenia: An update for the 21 st century. *Journal of Psychopharmacology*, *29*(2), 97–115. https://doi.org/10.1177/0269881114563634

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/72.991427

Hu, C., Ju, R., Shen, Y., Zhou, P., & Li, Q. (2016). Clinical decision support for Alzheimer's disease based on deep learning and brain network. In *2016 IEEE International Conference on Communications (ICC)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICC.2016.7510831

Huang, P., Xi, Y., Lu, Z.-L., Chen, Y., Li, X., Li, W., … Yin, H. (2015). Decreased bilateral thalamic gray matter volume in first-episode schizophrenia with prominent hallucinatory symptoms: A volumetric MRI study. *Scientific Reports*, *5*(1), 14505. https://doi.org/10.1038/srep14505

Huhtaniska, S., Jääskeläinen, E., Hirvonen, N., Remes, J., Murray, G. K., Veijola, J., … Miettunen, J. (2017). Long-term antipsychotic use and brain changes in schizophrenia - a systematic review and meta-analysis. *Human Psychopharmacology: Clinical and Experimental*, *32*(2), e2574. https://doi.org/10.1002/hup.2574

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., … others. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, *80*, 360–378.

Hutton, C., De Vita, E., Ashburner, J., Deichmann, R., & Turner, R. (2008). Voxel-based cortical thickness measurements in MRI. *NeuroImage*, *40*(4), 1701–1710. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811908000803?via%3Dihub#bib10

Hutton, C., Draganski, B., Ashburner, J., & Weiskopf, N. (2009). A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, *48*(2), 371–380. https://doi.org/10.1016/j.neuroimage.2009.06.043

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455–2465. https://doi.org/10.1017/S0033291716001367

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., … Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

IntHout, J., Ioannidis, J. P. A., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*(8), 860–869. https://doi.org/10.1016/J.JCLINEPI.2015.03.017

Jablensky, A. (2016). Psychiatric classifications: validity and utility. *World Psychiatry*, *15*(1), 26–31. https://doi.org/10.1002/wps.20284

Jain, A. K., Duin, P. W., & Jianchang Mao. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819

Janssen, R. J., Mourão-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 798–808. https://doi.org/10.1016/J.BPSC.2018.04.004

Jayakumar, P. N., Venkatasubramanian, G., Gangadhar, B. N., Janakiramaiah, N., & Keshavan, M. S. (2005). Optimized voxel-based morphometry of gray matter volume in first-episode, antipsychotic-naïve schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *29*(4), 587–591. https://doi.org/10.1016/J.PNPBP.2005.01.020

Jin, H., & Mosweu, I. (2017). The Societal Cost of Schizophrenia: A Systematic Review. *PharmacoEconomics*, *35*(1), 25–42. https://doi.org/10.1007/s40273-016-0444-6

Job, D. E., Whalley, H. C., McConnell, S., Glabus, M., Johnstone, E. C., & Lawrie, S. M. (2002). Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry. *NeuroImage*, *17*(2), 880–889. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12377162

Johnstone, E. C., Crow, T. J., Frith, C. D., Husband, J., & Kreel, L. (1976). Cerebral ventricular size and cognitive impairment in chronic schizophrenia. *Lancet (London, England)*, *2*(7992),

924–926. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/62160

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, Berlin.

Jones, S. E., Buchbinder, B. R., & Aharon, I. (2000). Three-dimensional mapping of cortical thickness using Laplace's equation. *Human Brain Mapping*, *11*(1), 12–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10997850

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joyce, E. M. (2018). Organic psychosis: The pathobiology and treatment of delusions. *CNS Neuroscience & Therapeutics*, *24*(7), 598–603.

Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., … Koutsouleris, N. (2017). Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biological Psychiatry*, *82*(5), 330–338. https://doi.org/10.1016/J.BIOPSYCH.2016.10.028

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., … Koutsouleris, N. (2015). Detecting Neuroimaging Biomarkers for Schizophrenia: A Meta-Analysis of Multivariate Pattern Recognition Studies. *Neuropsychopharmacology*, *40*(7), 1742–1751. https://doi.org/10.1038/npp.2015.22

Kambeitz-Ilankovic, L., Haas, S. S., Meisenzahl, E., Dwyer, D. B., Weiske, J., Peters, H., … Koutsouleris, N. (2019). Neurocognitive and neuroanatomical maturation in the clinical high-risk states for psychosis: A pattern recognition study. *NeuroImage: Clinical*, *21*, 101624. https://doi.org/10.1016/J.NICL.2018.101624

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179. https://doi.org/10.1038/mp.2012.105

Kellmeyer, P. (2018). Big Brain Data: On the Responsible Use of Brain Data from Clinical and Consumer-Directed Neurotechnological Devices. *Neuroethics*, 1–16.

Kellmeyer, P., Cochrane, T., Müller, O., Mitchell, C., Ball, T., Fins, J. J., & Biller-Andorno, N. (2016). The effects of closed-loop medical devices on the autonomy and accountability of persons and systems. *Cambridge Quarterly of Healthcare Ethics*, *25*(4), 623–633.

Kennedy, D. P., & Courchesne, E. (2008). The intrinsic functional organization of the brain is altered in autism. *NeuroImage*, *39*(4), 1877–1885.

https://doi.org/10.1016/j.neuroimage.2007.10.052

Keshavan, M. S., & Kaneko, Y. (2013). Secondary psychoses: an update. *World Psychiatry*, *12*(1), 4–15.

Keymer-Gausset, A., Alonso-Solís, A., Corripio, I., Sauras-Quetcuti, R. B., Pomarol-Clotet, E., Canales-Rodriguez, E. J., … Portella, M. J. (2018). Gray and white matter changes and their relation to illness trajectory in first episode psychosis. *European Neuropsychopharmacology*, *28*(3), 392–400. https://doi.org/10.1016/J.EURONEURO.2017.12.117

Khandanpour, N., Hoggard, N., & Connolly, D. J. A. (2013). The role of MRI and CT of the brain in first episodes of psychosis. *Clinical Radiology*, *68*(3), 245–250.

Kim, G.-W., Kim, Y.-H., & Jeong, G.-W. (2017). Whole brain volume changes and its correlation with clinical symptom severity in patients with schizophrenia: A DARTEL-based VBM study. *PLOS ONE*, *12*(5), e0177251. https://doi.org/10.1371/journal.pone.0177251

Kim, J.-J., Crespo-Facorro, B., Andreasen, N. C., O'Leary, D. S., Magnotta, V., & Nopoulos, P. (2003). Morphology of the lateral superior temporal gyrus in neuroleptic naïve patients with schizophrenia: relationship to symptoms. *Schizophrenia Research*, *60*(2–3), 173–181. https://doi.org/10.1016/S0920-9964(02)00299-2

Kim, J., Calhoun, V. D., Shim, E., & Lee, J.-H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, *124*, 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Kong, L., Herold, C. J., Zöllner, F., Salat, D. H., Lässer, M. M., Schmid, L. A., … Schröder, J. (2015). Comparison of grey matter volume and thickness for analysing cortical changes in chronic schizophrenia: A matter of surface area, grey/white matter intensity contrast, and curvature. *Psychiatry Research: Neuroimaging*, *231*(2), 176–183. https://doi.org/10.1016/J.PSCYCHRESNS.2014.12.004

Korver, N., Quee, P. J., Boos, H. B. M., Simons, C. J. P., & de Haan, L. (2012). Genetic Risk and Outcome of Psychosis (GROUP), a multi site longitudinal cohort study focused on gene-environment interaction: objectives, sample characteristics, recruitment and assessment methods. *International Journal of Methods in Psychiatric Research*, *21*(3), 205–221. https://doi.org/10.1002/mpr.1352

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nature Reviews Neuroscience*, *6*(9), 691–702. https://doi.org/10.1038/nrn1747

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Krogh, A., & Hertz, J. A. (1992). A Simple Weight Decay Can Improve Generalization. In D. S. Lippman, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems, vol. 4* (pp. 950–957). Morgan Kaufmann.

Kuang, D., Guo, X., An, X., Zhao, Y., & He, L. (2014). Discrimination of ADHD Based on fMRI Data with Deep Belief Network BT - Intelligent Computing in Bioinformatics. In D.-S. Huang, K. Han, & M. Gromiha (Eds.) (pp. 225–232). Cham: Springer International Publishing.

Kuang, D., & He, L. (2014). Classification on ADHD with deep learning. *Proceedings - 2014 International Conference on Cloud Computing and Big Data, CCBD 2014*, 27–32. https://doi.org/10.1109/CCBD.2014.42

Kumar, A., & Gopal, M. (2011). Reduced one-against-all method for multiclass SVM classification. *Expert Systems with Applications*, *38*(11), 14238–14248. https://doi.org/https://doi.org/10.1016/j.eswa.2011.04.237

Landhuis, E. (2017). Neuroscience: Big brain, big data. *Nature*, *541*(7638), 559–561. https://doi.org/10.1038/541559a

Langley, P. (2011). The changing science of machine learning. *Mach Learn*, *82*, 275–279. https://doi.org/10.1007/s10994-011-5242-y

Langley, P. (2016). *The Central Role of Cognition in Learning*. *Advances in Cognitive Systems* (Vol. 4). Retrieved from http://www.cogsys.org/papers/ACSvol4/paper2.pdf

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning* (pp. 473–480).

Lautrup, B., Hansen, L. K., Law, I., Morch, N., Svarer, C., & Strother, S. C. (1995). Massive weight sharing: A cure for extremely ill-posed problems. Paper presented at the proceedings of the workshop on supercomputing in brain research: From tomography to neural networks, Julich.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 8595–8598).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, H. W., Hong, S. B., Seo, D. W., Tae, W. S., & Hong, S. C. (2000). Mapping of functional organization in human visual cortex: electrical cortical stimulation. *Neurology*, *54*(4), 849–854. https://doi.org/10.1212/WNL.54.4.849

Lee, J. S., Park, H.-J., Chun, J. W., Seok, J.-H., Park, I.-H., Park, B., & Kim, J.-J. (2011). Neuroanatomical correlates of trait anhedonia in patients with schizophrenia: A voxel-based morphometric study. *Neuroscience Letters*, *489*(2), 110–114. https://doi.org/10.1016/J.NEULET.2010.11.076

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Li, F., Tran, L., Thung, K.-H., Ji, S., Shen, D., & Li, J. (2014). Robust deep learning for improved classification of AD/MCI patients. In *International Workshop on Machine Learning in Medical Imaging* (pp. 240–247).

Liao, J., Yan, H., Liu, Q., Yan, J., Zhang, L., Jiang, S., … Wang, F. (2015). Reduced paralimbic system gray matter volume in schizophrenia: Correlations with clinical variables, symptomatology and cognitive function. *Journal of Psychiatric Research*, *65*, 80–86. https://doi.org/10.1016/J.JPSYCHIRES.2015.04.008

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., … ADNI. (2015). Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140. https://doi.org/10.1109/TBME.2014.2372011

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early diagnosis of Alzheimer's

disease with deep learning. *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 1015–1018. https://doi.org/10.1109/ISBI.2014.6868045

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. D. (2015). Multi-Phase Feature Representation Learning for Neurodegenerative Disease Diagnosis, 1–10.

Luders, E., Narr, K. L., Thompson, P. M., Rex, D. E., Jancke, L., Steinmetz, H., & Toga, A. W. (2004). Gender differences in cortical complexity. *Nature Neuroscience*, *7*(8), 799–800. https://doi.org/10.1038/nn1277

Mahmoodi, J., Leckelt, M., van Zalk, M., Geukes, K., & Back, M. (2017). Big Data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, *18*, 57–62. https://doi.org/10.1016/J.COBEHA.2017.07.001

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry*, *80*(7), 552–561. Retrieved from https://www.sciencedirect.com/science/article/pii/S0006322316000020#f0020

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

McGorry, P. D., Killackey, E., & Yung, A. R. (2007). Early intervention in psychotic disorders: detection and treatment of the first episode and the critical early stages. *Medical Journal of Australia*, *187*(S7), S8–S10. https://doi.org/10.5694/j.1326-5377.2007.tb01327.x

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews*, *30*(1), 67–76. https://doi.org/10.1093/epirev/mxn001

McGuire, P. K., & Frith, C. D. (1996). Disordered functional connectivity in schizophrenia1. *Psychological Medicine*, *26*(04), 663. https://doi.org/10.1017/S0033291700037673

Mechelli, A., Prata, D., Kefford, C., & Kapur, S. (2015). Predicting clinical response in people at ultra-high risk of psychosis: a systematic and quantitative review. *Drug Discovery Today*, *20*(8), 924–927. https://doi.org/10.1016/j.drudis.2015.03.003

Mechelli, A., Price, C. J., Friston, K. J., & Ashburner, J. Voxel-Based Morphometry of the Human Brain: Methods and Applications, 1 Current Medical Imaging Reviews § (2005). https://doi.org/10.2174/1573405054038726

Meisenzahl, E. M., Koutsouleris, N., Bottlender, R., Scheuerecker, J., Jäger, M., Teipel, S. J., …

Möller, H.-J. (2008). Structural brain alterations at different stages of schizophrenia: A voxel-based morphometric study. *Schizophrenia Research*, *104*(1–3), 44–60. https://doi.org/10.1016/J.SCHRES.2008.06.023

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Meyer-Lindenberg, A. (2010). From maps to mechanisms through neuroimaging of schizophrenia. *Nature*, *468*(7321), 194–202. https://doi.org/10.1038/nature09569

Milham, M., Fair, D., Mennes, M., & Mostofsky, S. (2012). The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, *6*, 62. https://doi.org/10.3389/fnsys.2012.00062

Minzenberg, M. J., Laird, A. R., Thelen, S., Carter, C. S., & Glahn, D. C. (2009). Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Archives of General Psychiatry*, *66*(8), 811–822. https://doi.org/10.1001/archgenpsychiatry.2009.91

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870–877.

Modinos, G., Costafreda, S. G., van Tol, M.-J., McGuire, P. K., Aleman, A., & Allen, P. (2013). Neuroanatomy of auditory verbal hallucinations in schizophrenia: A quantitative meta-analysis of voxel-based morphometry studies. *Cortex*, *49*(4), 1046–1055. https://doi.org/10.1016/J.CORTEX.2012.01.009

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A. D. N., & others. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, *104*, 398–412. https://doi.org/10.1016/J.NEUROIMAGE.2014.10.002

Morris, S. E., & Cuthbert, B. N. (2012). Research Domain Criteria: cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience*, *14*(1), 29.

Mourão-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C. R., Shawe-Taylor, J., & Brammer, M. (2011). Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. *NeuroImage*, *58*(3), 793–804. https://doi.org/10.1016/J.NEUROIMAGE.2011.06.042

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., … Beckett, L. (2005a). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics*, *15*(4),

869–877. https://doi.org/10.1016/j.nic.2005.09.008

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., … Beckett, L. (2005b). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55–66. https://doi.org/10.1016/J.JALZ.2005.06.003

Mulders, P. C., van Eijndhoven, P. F., Schene, A. H., Beckmann, C. F., & Tendolkar, I. (2015). Resting-state functional connectivity in major depressive disorder: a review. *Neuroscience & Biobehavioral Reviews*, *56*, 330–344. https://doi.org/10.1016/j.neubiorev.2015.07.014

Munsell, B. C., Wee, C.-Y., Keller, S. S., Weber, B., Elger, C., da Silva, L. A. T., … Bonilha, L. (2015). Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *NeuroImage*, *118*, 219–230. https://doi.org/10.1016/j.neuroimage.2015.06.008

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., … Lopez, A. D. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2197–2223. https://doi.org/10.1016/S0140-6736(12)61689-4

Murray, G. K., Cheng, F., Clark, L., Barnett, J. H., Blackwell, A. D., Fletcher, P. C., … Jones, P. B. (2008). Reinforcement and Reversal Learning in First-Episode Psychosis. *Schizophrenia Bulletin*, *34*(5), 848–855. https://doi.org/10.1093/schbul/sbn078

Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244. https://doi.org/10.1007/s12021-013-9204-3.A

Nakamura, M., Nestor, P. G., Levitt, J. J., Cohen, A. S., Kawashima, T., Shenton, M. E., & McCarley, R. W. (2007). Orbitofrontal volume deficit in schizophrenia and thought disorder. *Brain*, *131*(1), 180–195. https://doi.org/10.1093/brain/awm265

Navari, S., & Dazzan, P. (2009). Do antipsychotic drugs affect brain structure? A systematic and critical review of MRI findings. *Psychological Medicine*, *39*(11), 1763–1777. https://doi.org/10.1017/S0033291709005315

Nekovarova, T., Fajnerova, I., Horacek, J., & Spaniel, F. (2014). Bridging disparate symptoms of schizophrenia: a triple network dysfunction theory. *Frontiers in Behavioral Neuroscience*, *8*, 171. https://doi.org/10.3389/fnbeh.2014.00171

Nielsen, R. E., Levander, S., Kjaersdam Telléus, G., Jensen, S. O. W., Østergaard Christensen, T., & Leucht, S. (2015). Second-generation antipsychotic effect on cognition in patients with schizophrenia-a meta-analysis of randomized clinical trials. *Acta Psychiatrica Scandinavica*, *131*(3), 185–196. https://doi.org/10.1111/acps.12374

Nieuwenhuis, M., Schnack, H. G., van Haren, N. E., Lappin, J., Morgan, C., Reinders, A. A., … Dazzan, P. (2017). Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *NeuroImage*, *145*, 246–253. https://doi.org/10.1016/j.neuroimage.2016.07.027

Nieuwenhuis, M., van Haren, N. E. M., Hulshoff Pol, H. E., Cahn, W., Kahn, R. S., & Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage*, *61*(3), 606–612. https://doi.org/10.1016/j.neuroimage.2012.03.079

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Nowlan, S. J., & Hinton, G. E. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation*, *4*(4), 473–493.

Nunes, A., Schnack, H. G., Ching, C. R. K., Agartz, I., Akudjedu, T. N., Alda, M., … Hajek, T. (2018). Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular Psychiatry*, 1. https://doi.org/10.1038/s41380-018-0228-9

O'Neill, A., Mechelli, A., & Bhattacharyya, S. (2018). Dysconnectivity of Large-Scale Functional Networks in Early Psychosis: A Meta-analysis. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sby094

Olabi, B., Ellison-Wright, I., McIntosh, A. M., Wood, S. J., Bullmore, E., & Lawrie, S. M. (2011). Are There Progressive Brain Changes in Schizophrenia? A Meta-Analysis of Structural Magnetic Resonance Imaging Studies. *Biological Psychiatry*, *70*(1), 88–96. https://doi.org/10.1016/J.BIOPSYCH.2011.01.032

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H.-U., Jönsson, B., CDBE2010 study group, & European Brain Council. (2012). The economic cost of brain disorders in Europe. *European Journal of Neurology*, *19*(1), 155–162. https://doi.org/10.1111/j.1468-1331.2011.03590.x

Olfson, M., Gerhard, T., Huang, C., Crystal, S., & Stroup, T. S. (2015). Premature Mortality Among Adults With Schizophrenia in the United States. *JAMA Psychiatry*, *72*(12), 1172. https://doi.org/10.1001/jamapsychiatry.2015.1737

Organization, W. H. (2004). *International statistical classification of diseases and related health problems* (Vol. 1). World Health Organization.

Organization, W. H. (2008). The global burden of disease: 2004 update.

Organization World Health. (1992). *International Classification of Diseases, Tenth Revision*. Geneva, Switzerland: World Health Organization.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140–1152. https://doi.org/10.1016/j.neubiorev.2012.01.004

Page, A., Turner, J. T., Mohsenin, T., & Oates, T. (2014). Comparing raw data and feature extraction for seizure detection with deep learning methods. In *The Twenty-Seventh International Flairs Conference*.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks, 1–9. Retrieved from http://arxiv.org/abs/1502.02506

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830. Retrieved from http://www.jmlr.org/papers/v12/pedregosa11a.html

Pelayo-Terán, J. M., Pérez-Iglesias, R., Ramírez-Bonilla, M., González-Blanch, C., Martínez-García, O., Pardo-García, G., … Crespo-Facorro, B. (2008). Epidemiological factors associated with treated incidence of first-episode non-affective psychosis in Cantabria: insights from the Clinical Programme on Early Phases of Psychosis. *Early Intervention in Psychiatry*, *2*(3), 178–187. https://doi.org/10.1111/j.1751-7893.2008.00074.x

Perälä, J., Suvisaari, J., Saarni, S. I., Kuoppasalmi, K., Isometsä, E., Pirkola, S., … Lönnqvist, J. (2007). Lifetime Prevalence of Psychotic and Bipolar I Disorders in a General Population. *Archives of General Psychiatry*, *64*(1), 19. https://doi.org/10.1001/archpsyc.64.1.19

Pereira, F., & Mitchell, T. (2008). Machine learning classifiers and fMRI : a tutorial overview, 1–

21.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Pettersson-Yeo, W., Allen, P., Benetti, S., McGuire, P., & Mechelli, A. (2011). Dysconnectivity in schizophrenia: Where are we now? *Neuroscience & Biobehavioral Reviews*, *35*(5), 1110–1124. https://doi.org/10.1016/J.NEUBIOREV.2010.11.004

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Dell'Acqua, F., Williams, S. C. R., Allen, P., … Mechelli, A. (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychological Medicine*, *43*(12), 2547–2562. https://doi.org/10.1017/S003329171300024X

Pina-Camacho, L., Del Rey-Mejias, A., Janssen, J., Bioque, M., Gonzalez-Pinto, A., Arango, C., … Parellada, M. (2016). Age at First Episode Modulates Diagnosis-Related Structural Brain Abnormalities in Psychosis. *Schizophrenia Bulletin*, *42*(2), 344–357. https://doi.org/10.1093/schbul/sbv128

Pinaya, W. H. L., Gadelha, A., Doyle, O. M., Noto, C., Zugman, A., Cordeiro, Q., … Sato, J. R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, *6*, 38897. https://doi.org/10.1038/srep38897

Pinaya, W. H. L., Mechelli, A., & Sato, J. R. (2018). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. https://doi.org/10.1002/hbm.24423

Plis, S. M., Amin, M. F., Chekroud, A., Hjelm, D., Damaraju, E., Lee, H. J., … Calhoun, V. D. (2018). Reading the (functional) writing on the (structural) wall: Multimodal fusion of brain structure and function via a deep neural network based translation approach reveals novel impairments in schizophrenia. *NeuroImage*, *181*, 734–747.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., … Calhoun, V. D. (2014). Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience*, *8*(August), 1–11. https://doi.org/10.3389/fnins.2014.00229

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, *17*(11), 1510–1517. https://doi.org/10.1038/nn.3818

Postema, M. C., Rooij, D. van, Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., … Francks, C. (2019). Altered structural brain asymmetry in autism spectrum disorder: large-scale analysis via the ENIGMA Consortium. *BioRxiv*, 570655. https://doi.org/10.1101/570655

Prata, D., Mechelli, A., & Kapur, S. (2014). Clinically meaningful biomarkers for psychosis: a systematic and quantitative review. *Neuroscience & Biobehavioral Reviews*, *45*, 134–141.

Prechelt, L. (1998). Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks, 11(4)*, 761e767. https://doi.org/10.1016/S0893-6080(98)00010-0.

Premkumar, P., Fannon, D., Sapara, A., Peters, E. R., Anilkumar, A. P., Simmons, A., … Kumari, V. (2015). Orbitofrontal cortex, emotional decision-making and response to cognitive behavioural therapy for psychosis. *Psychiatry Research: Neuroimaging*, *231*(3), 298–307. https://doi.org/10.1016/J.PSCYCHRESNS.2015.01.013

Puddephat, M. (2019). https://www.mikepuddephat.com/pages/149/3-magnetic-resonance-imaging.

Qiu, A., Gan, S. C., Wang, Y., & Sim, K. (2013). Amygdala–hippocampal shape and cortical thickness abnormalities in first-episode schizophrenia and mania. *Psychological Medicine*, *43*(07), 1353–1363. https://doi.org/10.1017/S0033291712002218

Radewicz, K., Garey, L. J., Gentleman, S. M., & Reynolds, R. (2000). Increase in HLA-DR Immunoreactive Microglia in Frontal and Temporal Cortex of Chronic Schizophrenics. *Journal of Neuropathology & Experimental Neurology*, *59*(2), 137–150. https://doi.org/10.1093/jnen/59.2.137

Radua, J., Borgwardt, S., Crescini, A., Mataix-Cols, D., Meyer-Lindenberg, A., McGuire, P. K., & Fusar-Poli, P. (2012). Multimodal meta-analysis of structural and functional brain changes in first episode psychosis and the effects of antipsychotic medication. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2325–2333. https://doi.org/10.1016/j.neubiorev.2012.07.012

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49. https://doi.org/10.1016/J.NEUROIMAGE.2017.01.066

Rapp, C., Bugra, H., Riecher-Rossler, A., Tamagni, C., & Borgwardt, S. (2012). Effects of cannabis use on human brain structure in psychosis: a systematic review combining in vivo

structural neuroimaging and post mortem studies. *Current Pharmaceutical Design*, *18*(32), 5070–5080.

Ren, W., Lui, S., Deng, W., Li, F., Li, M., Huang, X., … Gong, Q. (2013). Anatomical and Functional Brain Abnormalities in Drug-Naive First-Episode Schizophrenia. *American Journal of Psychiatry*, *170*(11), 1308–1316. https://doi.org/10.1176/appi.ajp.2013.12091148

Rimol, L. M., Nesvåg, R., Hagler, D. J., Bergmann, Ø., Fennema-Notestine, C., Hartberg, C. B., … Dale, A. M. (2012). Cortical Volume, Surface Area, and Thickness in Schizophrenia and Bipolar Disorder. *Biological Psychiatry*, *71*(6), 552–560. https://doi.org/10.1016/J.BIOPSYCH.2011.11.026

Roiz-Santiáñez, R., Pérez-Iglesias, R., Ortiz-García de la Foz, V., Tordesillas-Gutiérrez, D., Mata, I., Marco de Lucas, E., … Crespo-Facorro, B. (2011). Straight gyrus morphology in first-episode schizophrenia-spectrum patients. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *35*(1), 84–90. https://doi.org/10.1016/J.PNPBP.2010.09.002

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., … Davatzikos, C. (2018). Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. *Schizophrenia Bulletin*, *44*(5), 1035–1044. https://doi.org/10.1093/schbul/sbx137

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A Systematic Review of the Prevalence of Schizophrenia. *PLOS Medicine*, *2*(5), e141. https://doi.org/10.1371/journal.pmed.0020141

Salgado-Pineda, P., Baeza, I., Pérez-Gómez, M., Vendrell, P., Junqué, C., Bargalló, N., & Bernardo, M. (2003). Sustained attention impairment correlates to gray matter decreases in first episode neuroleptic-naive schizophrenic patients. *NeuroImage*, *19*(2), 365–375. https://doi.org/10.1016/S1053-8119(03)00094-6

Salvador, R., Radua, J., Canales-Rodríguez, E. J., Solanes, A., Sarró, S., Goikolea, J. M., … Pomarol-Clotet, E. (2017). Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *Plos One*, *12*(4), e0175683. https://doi.org/10.1371/journal.pone.0175683

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Sans-Sansa, B., McKenna, P. J., Canales-Rodríguez, E. J., Ortiz-Gil, J., López-Araquistain, L., Sarró, S., … Pomarol-Clotet, E. (2013). Association of formal thought disorder in schizophrenia with structural brain abnormalities in language-related cortical regions. *Schizophrenia Research*, *146*(1–3), 308–313. https://doi.org/10.1016/J.SCHRES.2013.02.032

Sapara, A., Cooke, M., Fannon, D., Francis, A., Buchanan, R. W., Anilkumar, A. P. P., … Kumari, V. (2007). Prefrontal cortex and insight in schizophrenia: A volumetric MRI study. *Schizophrenia Research*, *89*(1–3), 22–34. https://doi.org/10.1016/J.SCHRES.2006.09.016

Sarraf, S., & Tofighi, G. (2016). Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *ArXiv Preprint ArXiv:1603.08631*.

Sato, J. R., Rondina, J. M., & Mourão-Miranda, J. (2012). Measuring Abnormal Brains: Building Normative Rules in Neuroimaging Using One-Class Support Vector Machines. *Frontiers in Neuroscience*, *6*, 178. https://doi.org/10.3389/fnins.2012.00178

Savitz, J. B., Rauch, S. L., & Drevets, W. C. (2013). Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play. *Molecular Psychiatry*, *18*(5), 528.

Scanlon, C., Anderson-Schmidt, H., Kilmartin, L., McInerney, S., Kenney, J., McFarland, J., … McDonald, C. (2014). Cortical thinning and caudate abnormalities in first episode psychosis and their association with clinical outcome. *Schizophrenia Research*, *159*(1), 36–42. https://doi.org/10.1016/J.SCHRES.2014.07.030

Scarpazza, C., Tognin, S., Frisciata, S., Sartori, G., & Mechelli, A. (2015). False positive rates in Voxel-based Morphometry studies of the human brain: Should we be worried? *Neuroscience & Biobehavioral Reviews*, *52*, 49–55. https://doi.org/10.1016/j.neubiorev.2015.02.008

Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., … Veltman, D. J. (2017). Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Molecular Psychiatry*, *22*(6), 900–909. https://doi.org/10.1038/mp.2016.60

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schnack, H. G. (2017). Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases).

Schizophrenia Research. https://doi.org/10.1016/J.SCHRES.2017.10.023

Schnack, H. G., & Kahn, R. S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry*, *7*(MAR), 50. https://doi.org/10.3389/fpsyt.2016.00050

Schnack, H. G., Nieuwenhuis, M., van Haren, N. E. M. M., Abramovic, L., Scheewe, T. W., Brouwer, R. M., … Kahn, R. S. (2014). Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *NeuroImage*, *84*, 299–306. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811913009166

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience*, *10*(12), 885–892. https://doi.org/10.1038/nrn2753

Schultz, C. C., Fusar-Poli, P., Wagner, G., Koch, K., Schachtzabel, C., Gruber, O., … Schlösser, R. G. M. (2012). Multimodal functional and structural imaging investigations in psychosis research. *European Archives of Psychiatry and Clinical Neuroscience*, *262*(2), 97–106. https://doi.org/10.1007/s00406-012-0360-5

Schultz, C. C., Koch, K., Wagner, G., Roebel, M., Nenadic, I., Gaser, C., … Schlösser, R. G. M. (2010). Increased parahippocampal and lingual gyrification in first-episode schizophrenia. *Schizophrenia Research*, *123*(2–3), 137–144. https://doi.org/10.1016/J.SCHRES.2010.08.033

Schwarz, E., Doan, N. T., Pergola, G., Westlye, L. T., Kaufmann, T., Wolfers, T., … Meyer-Lindenberg, A. (2019). Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder. *Translational Psychiatry*, *9*(1), 12. https://doi.org/10.1038/s41398-018-0225-4

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11.

Shah, C., Zhang, W., Xiao, Y., Yao, L., Zhao, Y., Gao, X., … Lui, S. (2017). Common pattern of gray-matter abnormalities in drug-naive and medicated first-episode schizophrenia: a multimodal meta-analysis. *Psychological Medicine*, *47*(03), 401–413. https://doi.org/10.1017/S0033291716002683

Shen, X., Reus, L. M., Cox, S. R., Adams, M. J., Liewald, D. C., Bastin, M. E., … McIntosh, A. M.

(2017). Subcortical volume and white matter integrity abnormalities in major depressive disorder: findings from UK Biobank imaging data. *Scientific Reports*, *7*(1), 5547. https://doi.org/10.1038/s41598-017-05507-6

Shepherd, A. M., Matheson, S. L., Laurens, K. R., Carr, V. J., & Green, M. J. (2012). Systematic Meta-Analysis of Insula Volume in Schizophrenia. *Biological Psychiatry*, *72*(9), 775–784. https://doi.org/10.1016/J.BIOPSYCH.2012.04.020

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, *320*(21), 2199–2200.

Silverstein, S. M., & Keane, B. P. (2011). Perceptual Organization Impairment in Schizophrenia and Associated Brain Mechanisms: Review of Research from 2005 to 2010. *Schizophrenia Bulletin*, *37*(4), 690–699. https://doi.org/10.1093/schbul/sbr052

Simon, G. E., Stewart, C., Yarborough, B. J., Lynch, F., Coleman, K. J., Beck, A., … Hunkeler, E. M. (2018). Mortality Rates After the First Diagnosis of Psychotic Disorder in Adolescents and Young Adults. *JAMA Psychiatry*, *75*(3), 254. https://doi.org/10.1001/jamapsychiatry.2017.4437

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ArXiv Preprint ArXiv:1312.6034*.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Smieskova, R., Fusar-Poli, P., Allen, P., Bendfeldt, K., Stieglitz, R. D., Drewe, J., … Borgwardt, S. J. (2010). Neuroimaging predictors of transition to psychosis—A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *34*(8), 1207–1222. https://doi.org/10.1016/J.NEUBIOREV.2010.01.016

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). *SmoothGrad: removing noise by adding noise*. Retrieved from https://goo.gl/EfVzEE.

Smith, S. M., & Nichols, T. E. (2018). Statistical Challenges in “Big Data” Human Neuroimaging. *Neuron*, *97*(2), 263–268. https://doi.org/10.1016/J.NEURON.2017.12.018

Song, X., Quan, M., Lv, L., Li, X., Pang, L., Kennedy, D., … Fan, X. (2015). Decreased cortical thickness in drug naïve first episode schizophrenia: In relation to serum levels of BDNF.

Journal of Psychiatric Research, *60*, 22–28. https://doi.org/10.1016/J.JPSYCHIRES.2014.09.009

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for Simplicity: The All Convolutional Net.

Srinivasagopalan, S., Barry, J., Gurupur, V., & Thankachan, S. (2019). A deep learning approach for diagnosing schizophrenic patients. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–14.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stonnington, C. M., Chu, C., Klöppel, S., Jack, C. R., Ashburner, J., & Frackowiak, R. S. J. (2010). Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, *51*(4), 1405–1413. https://doi.org/10.1016/J.NEUROIMAGE.2010.03.051

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A Review of Reward Processing and Motivational Impairment in Schizophrenia. *Schizophrenia Bulletin*, *40*(Suppl 2), S107–S116. https://doi.org/10.1093/schbul/sbt197

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Suk, H.-I., Lee, S.-W., & Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, *101*, 569–582. https://doi.org/10.1016/J.NEUROIMAGE.2014.06.077

Suk, H.-I., Lee, S.-W., & Shen, D. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, *220*(2), 841–859. https://doi.org/10.1007/s00429-013-0687-3

Suk, H.-I., Lee, S.-W., & Shen, D. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure & Function*, *221*(5), 2569–2587. https://doi.org/10.1007/s00429-015-1059-y

Suk, H.-I., Wee, C.-Y., Lee, S.-W., & Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage*, *129*, 292–307.

Suk, H. Il, & Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8150 LNCS*(0 2), 583–590. https://doi.org/10.1007/978-3-642-40763-5_72

Sun, D., van Erp, T. G. M., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., … Cannon, T. D. (2009). Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms. *Biological Psychiatry*, *66*(11), 1055–1060. https://doi.org/10.1016/J.BIOPSYCH.2009.07.019

Surti, T. S., Corbera, S., Bell, M. D., & Wexler, B. E. (2011). Successful computer-based visual training specifically predicts visual memory enhancement over verbal memory improvement in schizophrenia. *Schizophrenia Research*, *132*(2–3), 131–134. https://doi.org/10.1016/j.schres.2011.06.031

Surti, T. S., & Wexler, B. E. (2012). A pilot and feasibility study of computer-based training for visual processing deficits in schizophrenia. *Schizophrenia Research*, *142*(1–3), 248–249. https://doi.org/10.1016/j.schres.2012.09.013

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.

Szendi, I., Kiss, M., Racsmány, M., Boda, K., Cimmer, C., Vörös, E., … Janka, Z. (2006). Correlations between clinical symptoms, working memory functions and structural brain abnormalities in men with schizophrenia. *Psychiatry Research: Neuroimaging*, *147*(1), 47–55. https://doi.org/10.1016/J.PSCYCHRESNS.2005.05.014

Takayanagi, Y., Kawasaki, Y., Nakamura, K., Takahashi, T., Orikabe, L., Toyoda, E., … Suzuki, M. (2010). Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *34*(1), 10–17. https://doi.org/10.1016/j.pnpbp.2009.09.004

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., … Suzuki, M. (2011). Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. *PLoS ONE*, *6*(6), 1–10. https://doi.org/10.1371/journal.pone.0021047

Tandon, N., & Tandon, R. (2018). Will Machine Learning Enable Us to Finally Cut the Gordian

Knot of Schizophrenia. *Schizophrenia Bulletin*, *44*(5), 939–941. https://doi.org/10.1093/schbul/sby101

Taylor, J. L., Blanton, R. E., Levitt, J. G., Caplan, R., Nobel, D., & Toga, A. W. (2005). Superior temporal gyrus differences in childhood-onset schizophrenia. *Schizophrenia Research*, *73*(2–3), 235–241. https://doi.org/10.1016/J.SCHRES.2004.07.023

Toga, A. W., Foster, I., Kesselman, C., Madduri, R., Chard, K., Deutsch, E. W., … Hood, L. (2015). Big Biomedical data as the key resource for discovery science. *Journal of the American Medical Informatics Association*, *22*(6), ocv077. https://doi.org/10.1093/jamia/ocv077

Tognin, S., Pettersson-Yeo, W., Valli, I., Hutton, C., Woolley, J., Allen, P., … Mechelli, A. (2013). Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis. *Frontiers in Psychiatry*, *4*, 187. https://doi.org/10.3389/fpsyt.2013.00187

Tordesillas-Gutierrez, D., Koutsouleris, N., Roiz-Santiañez, R., Meisenzahl, E., Ayesa-Arriola, R., Marco de Lucas, E., … Crespo-Facorro, B. (2015). Grey matter volume differences in non-affective psychosis and the effects of age of onset on grey matter volumes: A voxelwise study. *Schizophrenia Research*, *164*(1–3), 74–82. https://doi.org/10.1016/J.SCHRES.2015.01.032

Torres, U. S., Duran, F. L. S., Schaufelberger, M. S., Crippa, J. A. S., Louzã, M. R., Sallet, P. C., … Busatto, G. F. (2016). Patterns of regional gray matter loss at different stages of schizophrenia: A multisite, cross-sectional VBM study in first-episode and chronic illness. *NeuroImage: Clinical*, *12*, 1–15. https://doi.org/10.1016/J.NICL.2016.06.002

Ulloa, A. E., Plis, S., & Calhoun, V. D. Improving Classification Rate of Schizophrenia Using a Multimodal Multi-Layer Perceptron Model with Structural and Functional MRI, arXiv preprint § (2018). Retrieved from https://arxiv.org/pdf/1804.04591.pdf

Valli, I., Marquand, A. F., Mechelli, A., Raffin, M., Allen, P., Seal, M. L., & McGuire, P. (2016). Identifying Individuals at High Risk of Psychosis: Predictive Utility of Support Vector Machine using Structural and Functional MRI Data. *Frontiers in Psychiatry*, *7*, 52. https://doi.org/10.3389/fpsyt.2016.00052

van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., … Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*,

21(4), 547–553. https://doi.org/10.1038/mp.2015.63

van Erp, T. G. M., Preda, A., Nguyen, D., Faziola, L., Turner, J., Bustillo, J., … FBIRN. (2014). Converting positive and negative symptom scores between PANSS and SAPS/SANS. *Schizophrenia Research*, *152*(1), 289–294. https://doi.org/10.1016/J.SCHRES.2013.11.013

van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., … Turner, J. A. (2018). Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biological Psychiatry*, *84*(9), 644–654. https://doi.org/10.1016/J.BIOPSYCH.2018.04.023

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/J.NEUROIMAGE.2013.05.041

Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a “Big Data” science. *Brain Imaging and Behavior*, *8*(2), 323–331. https://doi.org/10.1007/s11682-013-9255-y

van Os, J., & Kapur, S. (2009). Schizophrenia. *Lancet (London, England)*, *374*(9690), 635–645. https://doi.org/10.1016/S0140-6736(09)60995-8

van Os, J., Kenis, G., & Rutten, B. P. F. (2010). The environment and schizophrenia. *Nature*, *468*(7321), 203–212. https://doi.org/10.1038/nature09563

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Varoquaux, G. (2017). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Venkatasubramanian, G., Jayakumar, P. N., Gangadhar, B. N., & Keshavan, M. S. (2008). Automated MRI parcellation study of regional volume and thickness of prefrontal cortex (PFC) in antipsychotic-naïve schizophrenia. *Acta Psychiatrica Scandinavica*, *117*(6), 420–431. https://doi.org/10.1111/j.1600-0447.2008.01198.x

Venkatasubramanian, Ganesan. (2010). Neuroanatomical correlates of psychopathology in antipsychotic-naïve schizophrenia. *Indian Journal of Psychiatry*, *52*(1), 28–36. https://doi.org/10.4103/0019-5545.58892

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the

neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58–75. https://doi.org/10.1016/J.NEUBIOREV.2017.01.002

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, *11*(Dec), 3371–3408. Retrieved from http://www.jmlr.org/papers/v11/vincent10a.html

Vita, A, De Peri, L., Deste, G., & Sacchetti, E. (2012). Progressive loss of cortical gray matter in schizophrenia: a meta-analysis and meta-regression of longitudinal MRI studies. *Translational Psychiatry*, *2*(11), e190. https://doi.org/10.1038/tp.2012.116

Vita, Antonio, De Peri, L., Deste, G., Barlati, S., & Sacchetti, E. (2015). The Effect of Antipsychotic Treatment on Cortical Gray Matter Changes in Schizophrenia: Does the Class Matter? A Meta-analysis and Meta-regression of Longitudinal Magnetic Resonance Imaging Studies. *Biological Psychiatry*, *78*(6), 403–412. https://doi.org/10.1016/J.BIOPSYCH.2015.02.008

Voets, N. L., Hough, M. G., Douaud, G., Matthews, P. M., James, A., Winmill, L., … Smith, S. (2008). Evidence for abnormalities of cortical development in adolescent-onset schizophrenia. *NeuroImage*, *43*(4), 665–675. https://doi.org/10.1016/j.neuroimage.2008.08.013

Walker, E. R., McGee, R. E., & Druss, B. G. (2015). Mortality in Mental Disorders and Global Disease Burden Implications. *JAMA Psychiatry*, *72*(4), 334. https://doi.org/10.1001/jamapsychiatry.2014.2502

Wang, F., Casalino, L. P., & Khullar, D. (2019). Deep Learning in Medicine—Promise, Progress, and Challenges. *JAMA Internal Medicine*, *179*(3), 293. https://doi.org/10.1001/jamainternmed.2018.7117

Wardenaar, K. J., & de Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Medicine*, *11*(1), 201. https://doi.org/10.1186/1741-7015-11-201

Wechsler, D. (1997). WAIS-III administration and scoring manual. 1997. *San Antonio, TX, The Psychological Corporation*.

Wegmayr, V., Aitharaju, S., & Buhmann, J. (2018). Classification of brain MRI with big data and deep 3D convolutional neural networks. In K. Mori & N. Petrick (Eds.), *Medical Imaging 2018: Computer-Aided Diagnosis* (p. 63). SPIE. https://doi.org/10.1117/12.2293719

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., … Trojanowski, J. Q. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, *13*(4), e1–e85. https://doi.org/10.1016/J.JALZ.2016.11.007

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization* (pp. 762–770). Berlin/Heidelberg: Springer-Verlag. https://doi.org/10.1007/BFb0006203

Willette, A. A., Calhoun, V. D., Egan, J. M., Kapogiannis, D., s Disease Neuroimaging Initiative, A., & others. (2014). Prognostic classification of mild cognitive impairment and Alzheimer′ s disease: MRI independent component analysis. *Psychiatry Research: Neuroimaging*, *224*(2), 81–88.

Wing, J. K., Babor, T., Brugha, T. S., Burke, J., Cooper, J. E., Giel, R., … Sartorius, N. (1990). Scan: Schedules fonr clinical assessment in neuropsychiatry. *Archives of General Psychiatry*, *47*(6), 589–593.

Winterburn, J. L., Voineskos, A. N., Devenyi, G. A., Plitman, E., de la Fuente-Sandoval, C., Bhagwat, N., … Chakravarty, M. M. (2017). Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophrenia Research*. https://doi.org/10.1016/J.SCHRES.2017.11.038

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*, *57*, 328–349. https://doi.org/10.1016/J.NEUBIOREV.2015.08.001

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., … Marquand, A. F. (2018). Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry*, *75*(11), 1146. Retrieved from https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2705762

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377. https://doi.org/10.1038/nn.4478

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified

statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, *4*(1), 58–73. https://doi.org/10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O

Wright, I. C., Rabe-Hesketh, S., Woodruff, P. W. R., David, A. S., Murray, R. M., & Bullmore, E. T. (2000). Meta-Analysis of Regional Brain Volumes in Schizophrenia. *American Journal of Psychiatry*, *157*(1), 16–25. https://doi.org/10.1176/ajp.157.1.16

Wylie, K. P., & Tregellas, J. R. (2010). The role of the insula in schizophrenia. *Schizophrenia Research*, *123*(2–3), 93–104. https://doi.org/10.1016/J.SCHRES.2010.08.027

Xiao, Y., Lui, S., Deng, W., Yao, L., Zhang, W., Li, S., … Gong, Q. (2015). Altered Cortical Thickness Related to Clinical Severity But Not the Untreated Disease Duration in Schizophrenia. *Schizophrenia Bulletin*, *41*(1), 201–210. https://doi.org/10.1093/schbul/sbt177

Xiao, Y., Yan, Z., Zhao, Y., Tao, B., Sun, H., Li, F., … Lui, S. (2017). Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*. https://doi.org/10.1016/J.SCHRES.2017.11.037

Xu, Y., Qin, W., Zhuo, C., Xu, L., Zhu, J., Liu, X., & Yu, C. (2017). Selective functional disconnection of the orbitofrontal subregions in schizophrenia. *Psychological Medicine*, *47*(09), 1637–1646. https://doi.org/10.1017/S0033291717000101

Yan, W., Plis, S., Calhoun, V. D., Liu, S., Jiang, R., Jiang, T.-Z., & Sui, J. (2017). Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP.2017.8168179

Yang, Z., Zhong, S., Carass, A., Ying, S. H., & Prince, J. L. (2014). Deep Learning for Cerebellar Ataxia Classification and Functional Score Regression. *Machine Learning in Medical Imaging. MLMI (Workshop)*, *8679*, 68–76. https://doi.org/10.1007/978-3-319-10581-9_9

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yassa, M., & Stark, C. (2009). A quantitative evaluation of cross-participant registration

techniques for MRI studies of the medial temporal lobe. *NeuroImage*, *44*(2), 319–327. https://doi.org/10.1016/j.neuroimage.2008.09.016

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *ArXiv Preprint ArXiv:1506.06579*.

Yung, A. R., Yuen, H. P., McGorry, P. D., Phillips, L. J., Kelly, D., Dell'Olio, M., … Buckby, J. (2005). Mapping the onset of psychosis: the Comprehensive Assessment of At-Risk Mental States. *The Australian and New Zealand Journal of Psychiatry*, *39*(11–12), 964–971. https://doi.org/10.1080/j.1440-1614.2005.01714.x

Yuste, R., Goering, S., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., … others. (2017). Four ethical priorities for neurotechnologies and AI. *Nature News*, *551*(7679), 159.

Zarogianni, E., Moorhead, T. W. J., & Lawrie, S. M. (2013). Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. *NeuroImage: Clinical*, *3*, 279–289. https://doi.org/10.1016/j.nicl.2013.09.003

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks BT - Computer Vision – ECCV 2014. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.) (pp. 818–833). Cham: Springer International Publishing.

Zeng, L.-L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., … Hu, D. (2018). Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. *EBioMedicine*, *30*, 74–85. https://doi.org/10.1016/J.EBIOM.2018.03.017

Zhang-James, Y., Helminen, E. C., Liu, J., ENIGMA-ADHD working group, T., Franke, B., Hoogman, M., & Faraone, S. V. (2019). Machine Learning Classification of Attention-Deficit/Hyperactivity Disorder Using Structural MRI Data Short running title: Image Classification for ADHD. *BioRxiv*. https://doi.org/10.1101/546671

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, L., Zheng, J., & McKeown, M. J. (2017). Deep learning based automatic diagnoses of attention deficit hyperactive disorder. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 962–966). IEEE. https://doi.org/10.1109/GlobalSIP.2017.8309103

Appendix 1. Publications derived from this thesis

Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Vieira S *et al* (2019). Neuroanatomical abnormalities in first-episode psychosis across independent samples: a multi-centre mega-analysis. *Psychological Medicine* 1–11. https://doi.org/10.1017/S0033291719003568

Received: 28 November 2018
Revised: 10 October 2019
Accepted: 21 November 2019

Key words:
First-episode psychosis; mega-analysis; multi-centre; neuroanatomy; voxel-based morphometry

Author for correspondence:
Qiyong Gong,
E-mail: qiyonggong@hmrrc.org.cn

CAMBRIDGE UNIVERSITY PRESS

Neuroanatomical abnormalities in first-episode psychosis across independent samples: a multi-centre mega-analysis

Sandra Vieira[1], Qiyong Gong[2,3,4], Cristina Scarpazza[1,5], Su Lui[2,3,4], Xiaoqi Huang[2,3,4], Benedicto Crespo-Facorro[6,7], Diana Tordesillas-Gutierrez[6,8], Víctor Ortiz-García de la Foz[6,7], Esther Setien-Suero[6,7], Floor Scheepers[9], Neeltje E.M. van Haren[9], René Kahn[9], Tiago Reis Marques[1], Simone Ciufolini[1], Marta Di Forti[10], Robin M Murray[1], Anthony David[11], Paola Dazzan[1], Philip McGuire[1] and Andrea Mechelli[1]

[1]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [2]Huaxi MR Research Center (HMRRC), Department of Radiology, West China Hospital of Sichuan University, Chengdu, China; [3]Psychoradiology Research Unit of Chinese Academy of Medical Sciences, West China Hospital of Sichuan University, Chengdu, Sichuan, China; [4]Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China; [5]Department of General Psychology, University of Padova, Padova, Italy; [6]CIBERSAM, Centro Investigación Biomédica en Red de Salud Mental, Madrid, Spain; [7]Department of Psychiatry, University Hospital Marqués de Valdecilla, School of Medicine, University of Cantabria-IDIVAL, Santander, Spain; [8]Neuroimaging Unit, Technological Facilities, Valdecilla Biomedical Research Institute IDIVAL, Santander, Cantabria, Spain; [9]Brain Centre Rudolf Magnus, University Medical Centre Utrecht, Utrecht, The Netherlands; [10]Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK and [11]UCL Institute of Mental Health, University College London, UK

Abstract

Background. Neuroanatomical abnormalities in first-episode psychosis (FEP) tend to be subtle and widespread. The vast majority of previous studies have used small samples, and therefore may have been underpowered. In addition, most studies have examined participants at a single research site, and therefore the results may be specific to the local sample investigated. Consequently, the findings reported in the existing literature are highly heterogeneous. This study aimed to overcome these issues by testing for neuroanatomical abnormalities in individuals with FEP that are expressed consistently across several independent samples.

Methods. Structural Magnetic Resonance Imaging data were acquired from a total of 572 FEP and 502 age and gender comparable healthy controls at five sites. Voxel-based morphometry was used to investigate differences in grey matter volume (GMV) between the two groups. Statistical inferences were made at $p < 0.05$ after family-wise error correction for multiple comparisons.

Results. FEP showed a widespread pattern of decreased GMV in fronto-temporal, insular and occipital regions bilaterally; these decreases were not dependent on anti-psychotic medication. The region with the most pronounced decrease – gyrus rectus – was negatively correlated with the severity of positive and negative symptoms.

Conclusions. This study identified a consistent pattern of fronto-temporal, insular and occipital abnormalities in five independent FEP samples; furthermore, the extent of these alterations is dependent on the severity of symptoms and duration of illness. This provides evidence for reliable neuroanatomical alternations in FEP, expressed above and beyond site-related differences in anti-psychotic medication, scanning parameters and recruitment criteria.

Introduction

Neuroanatomical abnormalities in schizophrenia have been well documented for the past four decades (Bora et al., 2011; Glahn et al., 2008). While the initial research was performed in patients with long-term schizophrenia (Ellison-Wright, Glahn, Laird, Thelen, & Bullmore, 2008), more recent studies have focussed on individuals in the early stages of the illness, when the effects of chronicity (Olabi et al., 2011; Vita, De Peri, Deste, & Sacchetti, 2012) and anti-psychotic medication (Radua et al., 2012; Shah et al., 2017; Vita, De Peri, Deste, Barlati, & Sacchetti, 2015) are minimal. The results of these studies, however, tend to be inconsistent from one investigation to another (Gao et al., 2018; Radua et al., 2012; Shah et al., 2017). For example, reports of insular abnormalities have been heterogeneous, with some studies reporting increased (Ren et al., 2013; Salgado-Pineda et al., 2003) and others decreased

(Chua et al., 2007; Jayakumar, Venkatasubramanian, Gangadhar, Janakiramaiah, & Keshavan, 2005; Venkatasubramanian, 2010) grey matter volume (GMV) in this region. A possible explanation for these inconsistencies is that most studies have used small sample sizes and therefore may have been under-powered. For example, in the most recent meta-analyses (Gao et al., 2018; Radua et al., 2012; Shah et al., 2017), out of a total of 37 studies included (after accounting for overlapping studies across meta-analyses), 20 had a total sample size of 60 or less. Studies with small sample sizes are likely to result in overestimates of effect size and low reproducibility due to low statistical power (Button et al., 2013); which suggests that some of these small studies may have suffered from an increased risk of false positives. In addition to being under-powered, different studies have also varied significantly in terms of their methods such as recruitment criteria, imaging acquisition parameters, pre-processing and statistical analysis (Radua et al., 2012). Furthermore, the vast majority of studies have examined participants from a single research site, raising the possibility that the results might be specific to the characteristics of the local sample investigated.

To overcome some of these limitations, the ENIGMA consortium developed a standardized pipeline detailing data pre-processing and analysis procedures; once data are analysed, single-site results are pooled and summarized in a meta-analysis. This approach has led to unprecedented sample sizes in schizophrenia research, with two recent studies of cortical abnormalities in 4474 patients and 5098 controls (van Erp et al., 2018), and subcortical changes in a smaller, albeit still impressive, sample of 2028 patients and 2540 controls (van Erp et al., 2016). However, although this approach mitigates some of the main limitations of the traditional meta-analysis by reducing the heterogeneity of the pooled single studies, findings still rely on reported results from individual studies, which may result in limited accuracy (Shah et al., 2017). Multi-centre mega-analyses, involving the pre-processing and integration of data from independent studies in one single statistical analysis, provide an opportunity to overcome this limitation. Gupta et al. (2014) analysed neuroanatomical abnormalities in the first mega-analysis in schizophrenia in a sample comprised of 784 individuals with established schizophrenia and 936 healthy controls (HC) collected from 23 sites. More recently, Rozycki et al. (2018) analysed data from five sites totalling 448 HC and 387 patients with chronic schizophrenia. Similar mega-analytic efforts focussed on the initial stages of the illness, when the effects of confounders are minimal, are still non-existent and evidence is still reliant on small-to-modest sized studies (Gao et al., 2018; Shah et al., 2017).

In light of the limitations of the existing literature, the aim of this study was to use a multi-centre mega-analytic approach to test for neuroanatomical changes in first-episode psychosis (FEP) that are consistent across independent samples. Based on the findings of the recent meta-analyses (Gao et al., 2018; Radua et al., 2012; Shah et al., 2017), we hypothesize that (i) patients would show GMV decrease in a distributed bilateral network including fronto-temporal and insular areas, consistently across the five independent samples; (ii) given the previous reports of symptom-dependent neuroanatomical alterations in psychosis (Fusar-Poli, Radua, McGuire, & Borgwardt, 2012; Tang et al., 2012), these decreases would be more pronounced in patients with more severe symptoms; and (iii) consistent with the existing evidence of progressive neuroanatomical changes in psychosis (Olabi et al., 2011; Vita et al., 2012), these decreases would be more pronounced in patients with longer duration of illness.

Methods

Subjects

A total of 1074 participants were included in the analysis. The total sample comprised of data collected from FEP patients and HC recruited as part of five independent studies, from four sites, all of which were previously published: Chengdu (China) (Gong et al., 2015), London (England) (GAP study; Di Forti et al., 2009), Santander (Spain) (PAFIP study; Pelayo-Terán et al., 2008) and Utrecht (The Netherlands) (GROUP study; Korver, Quee, Boos, Simons, & de Haan, 2012). Below is a description of the recruitment criteria for each study. All patients were experiencing their first psychotic episode, defined as the first manifestation of psychotic symptoms meeting criteria for a psychotic disorder, as specified by the DSM-IV (APA, 2000) or ICD-10 (WHO, 2004). For each of the five sites, ethical approval was granted from the relevant Ethics Committees, and written informed consent was obtained from all participants. Demographic and clinical data for patients and HC within each site are summarized in Table 1.

Site1: Chengdu, China

First-episode patients were recruited from the West China Hospital of Sichuan University, Chengdu, China as part of a wider study of psychiatric disorder diagnosis. Diagnosis of first episode of schizophrenia was determined by the consensus of two clinical psychiatrists using the Structured Interview for the DSM-IV Axis I Disorder (SCID) (First, Gibbon, Spitzer, & Williams, 1997). At the time of scanning, all patients were medication-naïve. HC were recruited by poster advertisement and screened using the SCID-I to confirm the lifetime absence of psychiatric disorders, as well as interviewed and subsequently excluded if they had any known history of psychiatric illness in first-degree relatives. Participants were excluded if they met any of the following criteria: (i) history of drug or alcohol abuse, (ii) pregnancy, and (iii) any physical illness such as hepatitis, cardiovascular disease or neurological disorder, as assessed by interview and review of medical records.

Site 2: London, England

Participants were recruited from the South London and Maudsley Foundation Trust and scanned at the Institute of Psychiatry, Psychology and Neuroscience. All patients meeting ICD-10 criteria for a diagnosis of psychosis (codes F20–F29 and F30–F33) (World Health Organization, 2004) were invited to participate in the study; patients with a diagnosis of organic psychosis were later excluded. HC were recruited through local advertisement from the same geographical areas as patients. A screening tool (Psychosis Screening Questionnaire; Bebbington & Nayani, 1995) was used to exclude the presence of psychotic symptomatology or a history of psychotic illness in controls. Additional exclusion criteria for all participants included learning disabilities (based as an IQ < 70), current or past neurological illness, brain injury with the loss of consciousness for more than 1 h and suspected or confirmed pregnancy.

Sites 3 and 4: Santander A and B, Spain

Data from Santander A and B were acquired as part of the same large prospective longitudinal study on the first non-affective episode psychosis in Cantabria, although with two different scanners. Individuals with FEP were recruited from both inpatient units and

Table 1. Demographic and clinical characteristics for FEP and HC for each site and total sample

		Chengdu, China (*N* = 240)		London, England (*N* = 168)		Santander A, Spain (*N* = 257)		Santander B, Spain (*N* = 223)		Utrecht, The Netherlands (*N* = 186)		Total (*N* = 1074)	
		HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP
N		118	122	92	76	113	144	78	145	101	85	502	572
Gender (%)	M	56 (48)	55 (45)	37 (40)	41(54)	70 (62)	81 (61)	48 (61)	89 (61)	69 (68)	68(80)	280 (56)	341 (60)
	F	62 (52)	67 (55)	55 (60)	35 (46)	43 (38)	56 (39)	30 (39)	56 (39)	32 (32)	17 (20)	222 (44)	231 (40)
		$\chi^2 = 0.1$, ns		$\chi^2 = 3.2$, ns		$\chi^2 = 0.0$, ns		$\chi^2 = 0.0$, ns		$\chi^2 = 3.2$, ns		$\chi^2 = 1.7$, ns	
Age M (s.d.)		25.8 (8.0)	27.0 (7.3)	26.5 (6.5)	27.0 (6.8)	29.7 (7.7)	29.3 (8.1)	28.0 (7.4)	29.5 (8.7)	26.8 (8.2)	25.4 (5.9)	27.8 (7.5)	27.7 (8.0)
		$t = 1.8$, ns		$t = 0.4$, ns		$t = 0.5$, ns		$t = 1.4$, ns		$t = 1.3$, ns		$t = 0.1$, ns	
TIV (L) M (s.d.)		1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.4 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)	1.5 (0.1)	1.5 (0.2)
		$t = 0.9$, ns		$t = 0.4$, ns		$t = 0.7$, ns		$t = 0.2$, ns		$t = 0.4$, ns		$t = 0.8$, ns	
Positive symptoms M (s.d.)		–	24.5 (6.9)[a]	–	13.7 (5.5)[a]	–	14.3 (4.4)[b]	–	13.5 (4.3)[b]	–	15.8 (6.3)[a]	–	–
Negative symptoms M (s.d.)		–	18.6 (8.6)[a]	–	15.7 (6.0)[a]	–	6.2 (5.0)[c]	–	6.2 (5.0)[c]	–	16.2 (6.9)[a]	–	–
Duration of illness (yrs) Med (IQR)		–	0.3 (0.9)	–	1.1 (1.3)	–	0.4 (0.7)	–	0.3 (1.0)	–	0.6 (1.4)	–	0.4 (0.9)
Antipsychotic medication (N) (naïve/typical/atypical/NA)		–	122/0/0/0	–	7/2/56/11	–	2/0/142/0	–	2/20/116/7	–	0/3/48/34	–	133/25/362/52

TIV, total intra-cranial volume; L, litres; M, male; F, female; FEP, first-episode psychosis; HC, healthy controls; Med, median; M, mean; s.d., standard deviation; NA, not available.
[a]PANSS: Positive and Negative Symptoms Scale.
[b]SAPS: Scale for the Assessment of Negative Symptoms.
[c]SANS: Scale for the Assessment of Negative Symptoms.
ns: $p > 0.05$.

community mental health care centres. Patients were included if they met the following criteria: (1) age 15–60 years; (2) DSM-IV criteria for a principal diagnosis of schizophrenia, schizophreniform disorder, schizoaffective disorder, brief reactive psychosis or not otherwise specified psychosis; and (3) no prior treatment with anti-psychotic medication or, if previously treated, a total life time of adequate anti-psychotic treatment of <6 weeks. Patients with DSM-IV diagnoses of mental retardation or substance dependence (except nicotine dependence) were excluded. Age- and gender-matched HC were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illnesses, including substance dependence and significant loss of consciousness, as determined by using an abbreviated version of the Comprehensive Assessment of Symptoms and History (CASH) (Andreasen, Flaum, & Arndt, 1992). The absence of psychosis in first-degree relatives was confirmed by clinical records and family interview.

Site 5: Utrecht, The Netherlands

Patients were identified through clinicians working in regional psychosis departments or academic centres and were included if they met the following criteria: (1) age range of 16–50 years; (2) a diagnosis of non-affective psychotic disorder according to the DSM-IV; (3) good command of the Dutch language; and (4) able and willing to give written informed consent. Controls were selected through a system of random mailings to addresses in the catchment areas of the cases and were included if the following criteria were met: (1) age range of 16 and 50 years, (2) no lifetime psychotic disorder, (3) no first-degree family member with a lifetime psychotic disorder, (4) good command of the Dutch language, and (5) able and willing to give written informed consent.

Magnetic resonance imaging (MRI) data acquisition

At all five sites, volumetric MRIs were acquired using a T1-weighted protocol. At three sites, the scanner field strength was 3T, and at two sites, it was 1.5T. The details of the MRI acquisition sequence for each site can be found in the online Supplementary material sTable 1.

Data analysis

Socio-demographic and clinical parameters

Differences between FEP and HC in gender, age and total intracranial volume (TIV) were assessed with a χ^2 and independent-sample t test for categorical and continuous data, respectively, using SPSS v24.

Pre-processing

From the initial pool of 1249 images made available, 21 were excluded due to scanner artefacts and gross anatomical abnormalities, 71 due to excessive noise and a further 83 were excluded to keep the maximum and minimum age (18–55) the same across all sites. Differences in GMV between HC and FEP were examined using voxel-based morphometry (VBM), as implemented in SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) running under MATLAB 9 (The MathWorks, Inc, Natick, Massachusetts, USA) (Ashburner & Friston, 2005). The following steps were followed for the pre-processing of each site: (1) checking for scanner artefacts and gross anatomical abnormalities for each subject; (2) setting the anterior commissure as the origin of the stereotaxic space and reorienting the image along the anterior commissure–posterior commissure (AC–PC) line; and (3) segmenting the image into grey matter, white matter and CSF maps. Next, all available images were used to create a study-specific template as implemented by the DARTEL (diffeomorphic anatomical registration using exponentiated lie algebra) toolbox (Ashburner, 2007). This procedure warps the grey matter and white matter partitions into a new study-specific reference space representing an average of all the subjects included in the analysis, thus maximizing accuracy and sensitivity (Yassa & Stark, 2009). Finally, GMV maps were normalized to the Montreal Neurological Institute (MNI) template and subsequently smoothed with an 8 mm Gaussian filter. A 'modulation step' was also included in the normalization step to preserve the information about the absolute grey matter values (Mechelli, Price, Friston, & Ashburner, 2005). The final smoothed, modulated, normalized data were used for the statistical analysis.

To assess the reliability of our findings, the analysis pipeline described above was replicated using: (1) CAT12 toolbox (http://dbm.neuro.uni-jena.de/cat/), (2) a template built with a homogeneous (equal number of patients and controls across sites) subsample and (3) different size kernels for smoothing. Results are shown in the online Supplementary materials.

Statistical analysis

Statistical analysis was carried out using an analysis of variance, with diagnostic group and scanning site as factors, resulting in 10 experimental groups. Age and gender were included as covariates of no interest. The option of proportional scaling was selected to remove confounding driven by global differences. Neuroanatomical alterations in patients with FEP relative to HC consistent across the five datasets were identified using the 'inclusive masking' option as implemented in SPM software. This option allowed us to test for voxels which showed (i) an overall statistically significant difference between patients and HC across all sites ($p < 0.05$ FWE corrected) and (ii) at least a strong trend at each site ($p < 0.05$ uncorrected). Specifically, this consisted of the following steps in SPM: (i) comparing all FEP against all HC at $p < 0.05$ FWE corrected using an overall main contrast – $FEP_{all\ sites}$ *v.* $HC_{all\ sites}$ (e.g. $FEP_{all\ sites} < HC_{all\ sites}$), (ii) overlaying this contrast with a second set of five FEP *v.* HC contrasts, one for each site (e.g. $FEP_{site\ 1} < HC_{site\ 1}$) at $p < 0.05$ uncorrected each, and finally (iii) identifying voxels of increased/decreased GMV in FEP relative to HC that survived both the overall and the site-level contrasts (Fig. 1). This procedure ensured that any overall statistically significant difference across the five sites would also be present at each site, at least at trend level. Statistical inferences were made using a minimum extent threshold of 50 voxels.

The TIV for each image was estimated by first calculating the volume of grey matter, white matter and CSF separately at each voxel from the segmented images; the total volume for each type of tissue was then calculated by summing the respective voxel-level volumes; finally, TIV was obtained by adding the volume of all three tissue types. The effects of symptom severity, illness duration and anti-psychotic medication on the identified clusters were estimated using Pearson's correlation between the values of GMV for the peak coordinate of each statistically significant cluster and each one of the clinical variables of interest. The raw psychotic symptom severity scores (acquired with either PANSS or SANS/SAPS) were first normalized to ensure

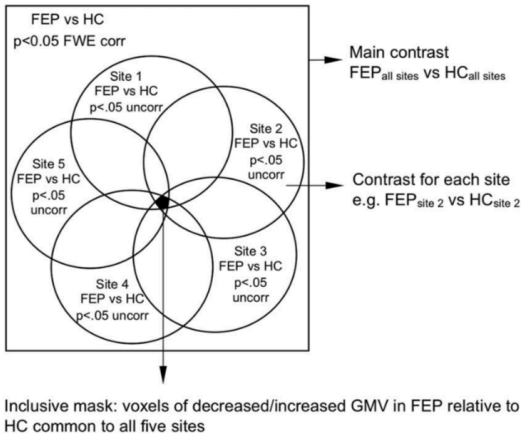


Fig. 1. Inclusive masking procedure used to identify neuroanatomical abnormalities in FEP relative to HC consistent across all five sites. Left: an overall contrast with all FEP against all HC ($p < 0.05$ FWE corrected) was combined with five site-level contrasts ($p < 0.05$ uncorrected); this allowed us to identify only the voxels that survived both types of contrasts (intersection of all contrasts in black).

comparability across sites. This normalization was achieved using the following formula:

$$\text{New score} = \frac{\text{Individual raw score} - \text{Minimum}}{\text{Maximum} - \text{Minimum}}$$

where Minimum and Maximum refer to the lowest and highest score allowed for either PANSS or SAPS/SANS. The resulting disease severity scores were scaled between 0 and 1. Across all sites (except site 1, where all patients were AP-naïve), AP medication dose was estimated by calculating the chlorpromazine equivalent (mg/day) for each individual according to Gardner, Murphy, O'Donnell, Centorrino, and Baldessarini (2010). Both chlorpromazine equivalent and duration of illness were log transformed. The statistical significance of Pearson's correlation was assessed using a p value<0.05 with Bonferroni correction for multiple comparisons.

Results

Socio-demographic and clinical parameters

There were no significant differences between FEP and HC in gender, age and TIV, both when considering all sites together and within each single site. Patients reported a comparable median duration of illness across sites (Table 1).

Decreased GMV in FEP compared to HC

Relative to HC, FEP showed a widespread pattern of decreased GMV in fronto-temporal, insular and occipital regions bilaterally (see Table 2 and Fig. 2*a*1). The most pronounced GMV decrease was found in the left gyrus rectus, located in the inferior frontal lobe (Fig. 2*a*2; the mean-plots for the remaining significant clusters are shown in the online Supplementary material sFig. 1); negative correlations were found between GVM in this region and severity of both positive and negative symptoms. The right lingual gyrus also showed negative correlations with both positive and negative symptoms as well as with the duration of illness. In addition, negative correlations were found between both the left inferior temporal gyrus and the left fusiform gyrus and positive symptoms (Table 3). No significant associations were detected between any brain region and anti-psychotic medication (Table 3). Scatter plots for the significant correlations are reported in the online Supplementary material sFig. 2.

Increased GMV in FEP compared to HC

A significant increase in GMV in FEP relative to HC was found in the right superior temporal gyrus (Table 2 and Fig. 2*b*). The volume of this region was not significantly associated with the severity of positive or negative symptoms, duration of illness or anti-psychotic medication (Table 3).

Discussion

Most previous studies on the neuroanatomical basis of FEP have used small samples recruited within a single site, and have yielded heterogeneous findings (Gao et al., 2018; Radua et al., 2012; Shah et al., 2017). The aim of this study was to use a multi-centre mega-analytic approach to identify neuroanatomical changes in FEP that are expressed consistently across several independent studies. As hypothesized, we found a widespread bilateral pattern of GMV decrease in fronto-temporal, insular and occipital regions. Some of these effects, particularly in the orbitofrontal and lingual gyri, were correlated with symptom severity and duration of illness. In addition, an increase in GMV was found in the right superior temporal lobe. Critically, all patients were experiencing their first episode of psychosis and one of the five samples was medication-naïve. In what follows, we discuss the brain structures that emerged from this study as well as their main role in the psychopathology of the early stages of psychosis.

Orbitofrontal cortex

A significant decrease in GMV was found in the two sub-regions of the orbitofrontal cortex (OFC), namely the gyrus rectus (straight gyrus) and the orbital gyrus (Buchanan et al., 2004; Nakamura et al., 2007). Grey matter deficits in the OFC have been reported in established psychosis (e.g. Kim, Kim, & Jeong, 2017; Kong et al., 2015; Rimol et al., 2012; Xu et al., 2017) and, to a lesser extent, in FEP (e.g. Crespo-Facorro, Kim, Andreasen, O'Leary, & Magnotta, 2000; Huang et al., 2015; Keymer-Gausset et al., 2018; Liao et al., 2015; Tordesillas-Gutierrez et al., 2015), consistent with the so-called 'hypo-frontality' hypothesis of psychosis; although increases in this region have also been observed (Gao et al., 2018). The OFC has been implicated in multiple functions, including cognitive flexibility, reward learning and decision making (see Kringelbach, 2005; Schoenbaum, Roesch, Stalnaker, & Takahashi, 2009 for a review), most of which are impaired in people with psychosis (Aas et al., 2014; Murray et al., 2008; Premkumar et al., 2015; Strauss, Waltz, & Gold, 2014). The gyrus rectus (straight gyrus) was the region with the most pronounced decrease in GMV within the OFC and the whole brain. Consistent with our finding, this region has been reported to be decreased in FEP regardless of anti-psychotic medication status in a recent meta-analysis (Shah et al., 2017). This is also consistent with the lack of a statistically significant association between this region and anti-psychotic medication found in the present study.

Table 2. MNI coordinates and *z* scores for regions showing GMV changes in FEP relative to the HC

Region	Peak MNI coordinates (*x*,*y*,*z*)	Cluster size (No of voxels)	*z*	*p*
Decreased GVM in FEP relative to HC				
L gyrus rectus	−6,34,−21	159	9.6	0.002
L med. orbital gyrus	−9,54,−15		8.4	
L sup. temporal pole	−21,8,−32	119	9.3	0.004
L fusiform gyrus	−20,2,−42		8.8	
R inf. temporal gyrus	56,−36,−20	607	9.1	<0.001
R mid. temporal gyrus	62,−32,−12		8.7	
L inf. temporal gyrus	−51,−52,−27	239	8.9	0.001
L fusiform gyrus	−46,−58, −22		8.6	
L mid. temporal gyrus	−58,−21,−14	161	8.8	0.002
R lingual gyrus	2,−80,−10	106	8.7	0.004
L mid. temporal gyrus	−52,−42,−18	63	8.5	0.009
L sup. temporal gyrus	−48, 18,−16	86	8.4	0.006
R insula	45,18,−8	88	8.3	0.006
Increased GVM in FEP relative to HC				
R sup. temporal gyrus	38,16,−38	338	8.1	<0.001

GMV, grey matter volume; FEP, first-episode psychosis; HC, healthy controls; L, left; R, right; med., medial; mid., middle; inf., inferior; sup., superior.

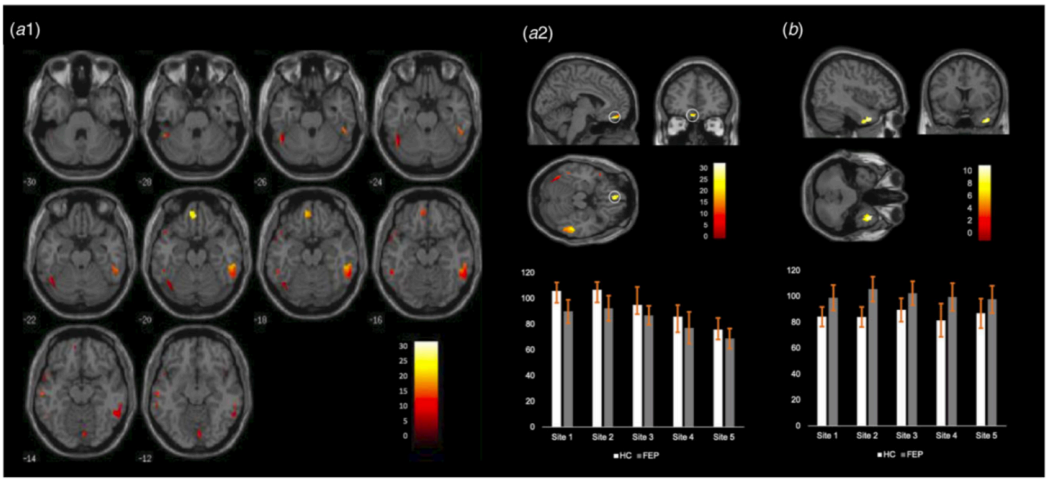


Fig. 2. (*a*1) Regions showing statistically significant decreases in FEP relative to HC across the whole brain. (*a*2) Top Location of the gyrus rectus (straight gyrus) where the most pronounced GMV decrease was found; bottom: mean and standard deviation of the GMV in this region for each site. (*b*) Top Location of the right superior temporal gyrus (the only region showing statistically significant GMV increase in FEP relative to HC); bottom: mean and standard deviation of the GMV in this region for each site.

A decrease in GMV in the gyrus rectus was also found in the largest single-site VBM study of first-episode patients to date which included 93 FEP participants and 175 controls (Meisenzahl et al., 2008); although evidence for normal volume has also been reported (Roiz-Santiáñez et al., 2011; Takayanagi et al., 2011). As hypothesized, GMV in the gyrus rectus was inversely related to positive and negative symptoms – consistent with previous studies (Kim et al., 2017; Sans-Sansa et al., 2013; Szendi et al., 2006).

Insula

Despite inconsistences across individual studies, most of the existing literature indicates deficits in the insular cortex of people with FEP, albeit with some inconsistencies in the exact location of the effect (Crespo-Facorro, Kim, Andreasen, O'Leary, Bockholt, et al., 2000; Gao et al., 2018; Shah et al., 2017). In the present investigation, it was the anterior part of the insula that showed reduced

Table 3. Pearson's correlations between regions showing GMV changes in FEP relative to the HC and symptom severity, illness duration and anti-psychotic medication

	Positive symptoms	Negative symptoms	Duration of illness	Anti-psychotic medication
Decreased GVM in FEP relative to HC				
L gyrus rectus	**−0.31**	**−0.20**	−0.10	−0.08
L med. orbital gyrus	**−0.17**	**−0.17**	−0.03	−0.06
L sup. temporal pole	−0.06	−0.07	−0.09	−0.04
L fusiform gyrus	−0.05	−0.05	−0.11	0.02
R inf. temporal gyrus	−0.11	−0.04	−0.08	−0.04
R mid. temporal gyrus	0.04	0.07	−0.01	−0.02
L inf. temporal gyrus	**−0.17**	−0.13	−0.10	−0.11
L fusiform gyrus	**−0.15**	−0.12	−0.08	−0.09
L mid. temporal gyrus	0.09	0.08	0.06	0.02
R lingual gyrus	**−0.20**	**−0.16**	**−0.18**	−0.13
L mid. temporal gyrus	−0.07	−0.06	−0.11	−0.08
L sup. temporal gyrus	0.02	−0.04	−0.06	−0.03
R insula	0.03	−0.02	**−0.15**	−0.08
Increased GVM in FEP relative to HC				
R sup. temporal gyrus	0.05	0.07	0.02	−0.08

GMV, grey matter volume; FEP, first-episode psychosis; HC, healthy controls; L, left; R, right; med., medial; mid., middle; inf., inferior; sup., superior.
Statistical inferences were made at $p < 0.05$ after Bonferroni correction for multiple comparisons based on the number of regions; this resulted in a p value of 0.05/14 = 0.0035. Statistically significant correlations are shown in bold.

GMV. This region plays an important role in salience processing (Menon & Uddin, 2010), emotional appraisal and social cognition (Eckert et al., 2009), all of which are affected in psychosis (Wylie & Tregellas, 2010). Notably, grey matter deficits in the insula, as well as in the gyrus rectus and superior temporal gyrus, have also been found in individuals at ultra-high risk for psychosis who later transitioned to psychosis (Smieskova et al., 2010); this suggests reduced GMV in this region may represent a neuroanatomical signature of vulnerability to psychosis rather than a marker of the actual illness. Furthermore, a GMV decrease in this region has been found to be expressed above and beyond ethnic variations in incidence and clinical expression (Gong et al., 2015).

Temporal cortex

Decreased GMV in temporal regions are amongst the most replicated findings in psychosis, including in FEP (Chan, Di, McAlonan, & Gong, 2011; Radua et al., 2012; Shah et al., 2017). In this study, several temporal regions showed GMV deficits, namely the superior, middle and inferior gyri as well as the temporal portion of the fusiform gyrus bilaterally. GMV deficits in the left superior temporal gyrus are thought to play a central role in auditory verbal hallucinations in FEP patients (Benetti et al., 2015; Modinos et al., 2013), possibly due to the role of this region in language perception and processing; it has been suggested that impairment to this region may lead to a misattribution of internal speech (Frith & Done 1988; Mechelli et al., 2007). The fusiform gyrus is also thought to play an important role in the psychopathology of psychosis, mainly due to its contribution to facial recognition (Haxby, Hoffman, & Gobbini, 2000, 2002), which is impaired in psychosis (see Green, Horan, & Lee, 2015; Barkl, Lah, Harris, & Williams, 2014 for a review) and is often seen as a proxy for the social cognition deficits characteristic of the illness (Green et al., 2015). Perhaps more challenging to interpret is the significant increase in GMV in the right superior temporal gyrus. Nevertheless, increases in patients relative to controls across the brain, including the temporal cortex, have been reported before (Kim et al., 2003; Lee et al., 2011; Radewicz, Garey, Gentleman, & Reynolds, 2000; Taylor et al., 2005), and are typically interpreted in terms of a 'compensatory mechanism' (Guo, Palaniyappan, Liddle, & Feng, 2016) or a transient inflammation resulting from increased apoptotic activity during which apoptotic cells are removed (Adler, Levine, DelBello, & Strakowski, 2005; Berger, Wood, & McGorry, 2003).

Lingual gyrus

Evidence supporting structural abnormalities in the lingual gyrus in FEP has not been as consistent, with some studies reporting decreased (Ellison-Wright et al., 2008) and others increased (Gao et al., 2018) GMV. Such inconsistency may be explained by medication status, as shown by Shah et al. (2017), where GMV of the lingual gyrus was decreased in anti-psychotic naive FEP patients but increased in FEP patients undergoing anti-psychotic treatment. However, in our study, which included both samples with and without exposure to anti-psychotics, there was a consistent GMV decrease in the lingual gyrus in each of the five sites, suggesting that a GMV decrease in this region may be present above and beyond medication status. Nevertheless, the lingual gyrus was significantly associated with anti-psychotic medication, therefore indicating that this region may be particularly prone to alterations when exposed to medication. The lingual gyrus is involved mainly in visual processing

(Hahn, Ross, & Stein, 2006; Lee, Hong, Seo, Tae, & Hong, 2000) which has been shown to be impaired in psychosis (see Butler, Silverstein, & Dakin, 2008; Silverstein & Keane, 2011 for a review) and are also thought to underlie some of the cognitive impairments characteristic of the illness (Contreras, Tan, Lee, Castle, & Rossell, 2018; Surti & Wexler, 2012; Surti, Corbera, Bell, & Wexler, 2011). The lingual gyrus also contributes to the evaluation of emotional faces (Fusar-Poli et al., 2009) which, together with the deficits found in the fusiform gyrus, may explain social cognition impairments in psychosis (Green et al., 2015).

Limitations

A first limitation of this study was that clinical data were acquired using different instruments (positive symptoms were assessed with either the PANSS or SAPS and negative symptoms with the PANSS or SANS). We overcame this limitation by normalizing individual scores within each scale as in the previous studies (Gong et al., 2019). The resulting scores were highly correlated ($r = 0.87$) with automated methods to convert scores between these two widely used scales (van Erp et al., 2014). A further limitation is that there were differences in age, gender and clinical presentation across the five samples. However, we do not think this undermines the reliability of our findings, since our statistical analysis tested for common neuroanatomical abnormalities across the five sites rather than site-specific effects. Additionally, the MRI data were not harmonized across sites using dedicated approaches such as ComBat (Johnson, Li, & Rabinovic, 2006), which could have improved the reliability of the results. The majority of VBM studies so far, including the present study, have used the DARTEL approach in-built in SPM to create study-specific templates. Although this is a well-established method, future studies could benefit from the use of recent alternative approaches, such as ANTs (http://stnava.github.io/ANTs/). A final limitation is that the five datasets differed with respect to anti-psychotic medication, with one sample being medication-naïve and the remaining four samples receiving various degrees of medication (Table 1). Critically, when we examined the impact of anti-psychotic medication on the findings, we found little evidence of statistically significant effects. This can be explained by the fact that our findings were based on the consistent neuroanatomical abnormalities across the five datasets, which included both medicated and non-medicated samples.

Conclusion

This study aimed to overcome the limitations of small and single-site studies by conducting a multi-centre mega-analysis of neuroanatomical abnormalities in FEP. To the best of our knowledge, this is the largest VBM study in FEP to date. We found that a widespread pattern of fronto-temporal, insular and occipital decreased GMV in FEP that were expressed consistently across five independent studies; overall, these decreases were not affected by anti-psychotic medication. This provides evidence for reliable neuroanatomical alternations in FEP, expressed above and beyond site-related differences in anti-psychotic medication, scanning parameters and recruitment criteria. With the increasing availability of larger datasets, future multi-centre mega-analyses could investigate the diagnostic specificity of these findings by integrating the data collected from people with different psychiatric diagnoses (Ellison-Wright & Bullmore, 2010; Gong et al., 2019).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291719003568

Financial support. This study was supported by the European Commission (PSYSCAN – Translating neuroimaging findings from research into clinical practice) (P.M., grant number 603196); International Cooperation and Exchange of the National Natural Science Foundation of China (Q.G. and A.M., grant numbers 81220108013, 8122010801, 81621003, 81761128023 and 81227002); Wellcome Trusts Innovator Award (A.M., grant number 208519/Z/17/Z) Italian Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) (C.R, grant number art.1, commi 314-337 legge 232/2016) and the Foundation for Science and Technology (FCT) (S.V., grant number SFRH/BD/103907/2014).

Conflict of interest. None.

References

Aas, M., Dazzan, P., Mondelli, V., Melle, I., Murray, R. M., & Pariante, C. M. (2014). A systematic review of cognitive function in first-episode psychosis, including a discussion on childhood trauma, stress, and inflammation. *Frontiers in Psychiatry*, *4*, 182. doi: 10.3389/fpsyt.2013.00182

Adler, C. M., Levine, A. D., DelBello, M. P., & Strakowski, S. M. (2005). Changes in gray matter volume in patients with bipolar disorder. *Biological Psychiatry*, *58*(2), 151–157. doi: 10.1016/J.BIOPSYCH.2005.03.022

Andreasen, N. C., Flaum, M., & Arndt, S. (1992). The Comprehensive Assessment of Symptoms and History (CASH). *Archives of General Psychiatry*, *49*(8), 615. doi: 10.1001/archpsyc.1992.01820080023004

APA (2000). Diagnostic and statistical manual of mental disorders 4th edition (DSM-IV-TR). Washington, DC: American Psychiatric Association.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Barkl, S. J., Lah, S., Harris, A. W. F., & Williams, L. M. (2014). Facial emotion identification in early-onset and first-episode psychosis: A systematic review with meta-analysis. *Schizophrenia Research*, *159*(1), 62–69. doi: 10.1016/J.SCHRES.2014.07.049

Bebbington, P., & Nayani, T. (1995). The psychosis screening questionnaire. *International Journal of Methods in Psychiatric Research*, *5*, 11–19.

Benetti, S., Pettersson-Yeo, W., Allen, P., Catani, M., Williams, S., Barsaglini, A., … Mechelli, A. (2015). Auditory verbal hallucinations and brain disconnectivity in the Perisylvian language network: A multimodal investigation. *Schizophrenia Bulletin*, *41*(1), 192–200. doi: 10.1093/schbul/sbt172

Berger, G. E., Wood, S., & McGorry, P. D. (2003). Incipient neurovulnerability and neuroprotection in early psychosis. *Psychopharmacology Bulletin*, *37*(2), 79–101. Retrieved September 6, 2018, from http://www.ncbi.nlm.nih.gov/pubmed/14566217.

Bora, E., Fornito, A., Radua, J., Walterfang, M., Seal, M., Wood, S. J. *et al.* (2011). Neuroanatomical abnormalities in schizophrenia: A multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophrenia Research*, *127*(1–3), 46–57. doi: 10.1016/J.SCHRES.2010.12.020

Buchanan, R. W., Francis, A., Arango, C., Miller, K., Lefkowitz, D. M., McMahon, R. P., … Pearlson, G. D. (2004). Morphometric assessment of the heteromodal association cortex in schizophrenia. *American Journal of Psychiatry*, *161*(2), 322–331. doi: 10.1176/appi.ajp.161.2.322

Butler, P. D., Silverstein, S. M., & Dakin, S. C. (2008). Visual perception and its impairment in schizophrenia. *Biological Psychiatry*, *64*(1), 40–47. doi: 10.1016/J.BIOPSYCH.2008.03.023

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Chan, R. C. K., Di, X., McAlonan, G. M., & Gong, Q. (2011). Brain anatomical abnormalities in high-risk individuals, first-episode, and chronic schizophrenia: An activation likelihood estimation meta-analysis of illness progression. *Schizophrenia Bulletin*, *37*(1), 177–188. doi: 10.1093/schbul/sbp073

Chua, S. E., Cheung, C., Cheung, V., Tsang, J. T. K., Chen, E. Y. H., Wong, J. C. H., … McAlonan, G. M. (2007). Cerebral grey, white matter and CSF in never-medicated, first-episode schizophrenia. *Schizophrenia Research*, *89*(1–3), 12–21. doi: 10.1016/j.schres.2006.09.009

Contreras, N. A., Tan, E. J., Lee, S. J., Castle, D. J., & Rossell, S. L. (2018). Using visual processing training to enhance standard cognitive remediation outcomes in schizophrenia: A pilot study. *Psychiatry Research*, *262*, 494–499. doi: 10.1016/J.PSYCHRES.2017.09.031

Crespo-Facorro, B., Kim, J.-J., Andreasen, N. C., O'Leary, D. S., Bockholt, H. J., & Magnotta, V. (2000). Insular cortex abnormalities in schizophrenia: A structural magnetic resonance imaging study of first-episode patients. *Schizophrenia Research*, *46*(1), 35–43.

Crespo-Facorro, B., Kim, J.-J., Andreasen, N. C., O'Leary, D. S., & Magnotta, V. (2000). Regional frontal abnormalities in schizophrenia: A quantitative gray matter volume and cortical surface size study. *Biological Psychiatry*, *48*(2), 110–119.

Di Forti, M., Morgan, C., Dazzan, P., Pariante, C., Mondelli, V., Marques, T. R., … Murray, R. M. (2009). High-potency cannabis and the risk of psychosis. *British Journal of Psychiatry*, *195*(06), 488–491. doi: 10.1192/bjp.bp.109.064220

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*(8), 2530–2541. doi: 10.1002/hbm.20688

Ellison-Wright, I., & Bullmore, E. (2010). Anatomy of bipolar disorder and schizophrenia: A meta-analysis. *Schizophrenia Research*, *117*(1), 1–12. doi: 10.1016/J.SCHRES.2009.12.022

Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M., & Bullmore, E. (2008). The anatomy of first-episode and chronic schizophrenia: An anatomical likelihood estimation meta-analysis. *American Journal of Psychiatry*, *165*(8), 1015–1023. doi: 10.1176/appi.ajp.2008.07101562

First, M. B., Gibbon, M., Spitzer, R. L., & Williams, J. (1997). *Structured clinical interview for DSM-IV axis II personality disorders*. Washington: American Psychiatric Press.

Frith, C. D., & Done, D. J. (1988). Towards a neuropsychology of schizophrenia. The British Journal of Psychiatry, *153*(4), 437–443.

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., … Politi, P. (2009). Functional atlas of emotional faces processing: A voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience: JPN*, *34*(6), 418–432. Retrieved September 5, 2018, from http://www.ncbi.nlm.nih.gov/pubmed/19949718.

Fusar-Poli, P., Radua, J., McGuire, P., & Borgwardt, S. (2012). Neuroanatomical maps of psychosis onset: Voxel-wise meta-analysis of antipsychotic-naive VBM studies. *Schizophrenia Bulletin*, *38*(6), 1297–1307. doi: 10.1093/schbul/sbr134

Gao, X., Zhang, W., Yao, L., Xiao, Y., Liu, L., Liu, J., … Lui, S. (2018). Association between structural and functional brain alterations in drug-free patients with schizophrenia: A multimodal meta-analysis. *Journal of Psychiatry & Neuroscience*, *43*(2), 131–142. doi: 10.1503/jpn.160219

Gardner, D. M., Murphy, A. L., O'Donnell, H., Centorrino, F., & Baldessarini, R. J. (2010). International consensus study of antipsychotic dosing. *American Journal of Psychiatry*, *167*(6), 686–693. doi: 10.1176/appi.ajp.2009.09060802

Glahn, D. C., Laird, A. R., Ellison-Wright, I., Thelen, S. M., Robinson, J. L., Lancaster, J. L., … Fox, P. T. (2008). Meta-analysis of gray matter anomalies in schizophrenia: Application of anatomic likelihood estimation and network analysis. *Biological Psychiatry*, *64*(9), 774–781. doi: 10.1016/J.BIOPSYCH.2008.03.031

Gong, Q., Dazzan, P., Scarpazza, C., Kasai, K., Hu, X., Marques, T. R., … Mechelli, A. (2015). A neuroanatomical signature for schizophrenia across different ethnic groups. *Schizophrenia Bulletin*, *41*(6), 1266–1275. doi: 10.1093/schbul/sbv109

Gong, Q., Scarpazza, C., Dai, J., He, M., Xu, X., Shi, Y., … Mechelli, A. (2019). A transdiagnostic neuroanatomical signature of psychiatric illness. *Neuropsychopharmacology*, *44*(5), 869.

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620–631. doi: 10.1038/nrn4005

Guo, S., Palaniyappan, L., Liddle, P. F., & Feng, J. (2016). Dynamic cerebral reorganization in the pathophysiology of schizophrenia: A MRI-derived cortical thickness study. *Psychological Medicine*, *46*(10), 2201–2214. doi: 10.1017/S0033291716000994

Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., … Buitelaar, J. (2014). Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. *Schizophrenia Bulletin*, *41*(5), 1133–1142. doi: 10.1093/schbul/sbu177.

Hahn, B., Ross, T. J., & Stein, E. A. (2006). Neuroanatomical dissociation between bottom–up and top–down processes of visuospatial selective attention. *NeuroImage*, *32*(2), 842–853. doi: 10.1016/j.neuroimage.2006.04.177

Haxby, J. V, Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*(1), 59–67.

Huang, P., Xi, Y., Lu, Z.-L., Chen, Y., Li, X., Li, W., … Yin, H. (2015). Decreased bilateral thalamic gray matter volume in first-episode schizophrenia with prominent hallucinatory symptoms: A volumetric MRI study. *Scientific Reports*, *5*(1), 14505. doi: 10.1038/srep14505

Jayakumar, P. N., Venkatasubramanian, G., Gangadhar, B. N., Janakiramaiah, N., & Keshavan, M. S. (2005). Optimized voxel-based morphometry of gray matter volume in first-episode, antipsychotic-naïve schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *29*(4), 587–591. doi: 10.1016/J.PNPBP.2005.01.020

Johnson, W. E., Li, C., & Rabinovic, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, *8*(1), 118–127.

Keymer-Gausset, A., Alonso-Solís, A., Corripio, I., Sauras-Quetcuti, R. B., Pomarol-Clotet, E., Canales-Rodriguez, E. J., … Portella, M. J. (2018). Gray and white matter changes and their relation to illness trajectory in first episode psychosis. *European Neuropsychopharmacology*, *28*(3), 392–400. doi: 10.1016/J.EURONEURO.2017.12.117

Kim, G.-W., Kim, Y.-H., & Jeong, G.-W. (2017). Whole brain volume changes and its correlation with clinical symptom severity in patients with schizophrenia: A DARTEL-based VBM study. *PLoS ONE*. Edited by K. Hashimoto. Public Library of Science, *12*(5), e0177251. doi: 10.1371/journal.pone.0177251.

Kim, J.-J., Crespo-Facorro, B., Andreasen, N. C., O'Leary, D. S., Magnotta, V., & Nopoulos, P. (2003). Morphology of the lateral superior temporal gyrus in neuroleptic naïve patients with schizophrenia: Relationship to symptoms. *Schizophrenia Research*, *60*(2–3), 173–181. doi: 10.1016/S0920-9964(02)00299-2

Kong, L., Herold, C. J., Zöllner, F., Salat, D. H., Lässer, M. M., Schmid, L. A., … Schröder, J. (2015). Comparison of grey matter volume and thickness for analysing cortical changes in chronic schizophrenia: A matter of surface area, grey/white matter intensity contrast, and curvature. *Psychiatry Research: Neuroimaging*, *231*(2), 176–183. doi: 10.1016/J.PSCYCHRESNS.2014.12.004

Korver, N., Quee, P. J., Boos, H. B. M., Simons, C. J. P., & de Haan, L. (2012). Genetic Risk and Outcome of Psychosis (GROUP), a multi site longitudinal cohort study focused on gene-environment interaction: Objectives, sample characteristics, recruitment and assessment methods. *International Journal of Methods in Psychiatric Research*, *21*(3), 205–221. doi: 10.1002/mpr.1352

Kringelbach, M. L. (2005). The human orbitofrontal cortex: Linking reward to hedonic experience. *Nature Reviews Neuroscience*, *6*(9), 691–702. doi: 10.1038/nrn1747

Lee, H. W., Hong, S. B., Seo, D. W., Tae, W. S., & Hong, S. C. (2000). Mapping of functional organization in human visual cortex: Electrical cortical stimulation. *Neurology*. Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology, *54*(4), 849–854. doi: 10.1212/WNL.54.4.849.

Lee, J. S., Park, H.-J., Chun, J. W., Seok, J.-H., Park, I.-H., Park, B., & Kim, J.-J. (2011). Neuroanatomical correlates of trait anhedonia in patients with schizophrenia: A voxel-based morphometric study. *Neuroscience Letters*, *489*(2), 110–114. doi: 10.1016/J.NEULET.2010.11.076

Liao, J., Yan, H., Liu, Q., Yan, J., Zhang, L., Jiang, S., … Wang, F. (2015). Reduced paralimbic system gray matter volume in schizophrenia: Correlations with

clinical variables, symptomatology and cognitive function. *Journal of Psychiatric Research*, *65*, 80–86. doi: 10.1016/J.JPSYCHIRES.2015.04.008

Mechelli, A., Allen, P., Amaro Jr, E., Fu, C. H., Williams, S. C., Brammer, M. J., … McGuire, P. K. (2007). Misattribution of speech and impaired connectivity in patients with auditory verbal hallucinations. *Human Brain Mapping*, *28*(11), 1213–1222.

Mechelli, A., Price, C. J., Friston, K. J., & Ashburner, J. (2005). *Voxel-Based Morphometry of the Human Brain: Methods and Applications*, *Current Medical Imaging Reviews*. doi: 10.2174/1573405054038726.

Meisenzahl, E. M., Koutsouleris, N., Bottlender, R., Scheuerecker, J., Jäger, M., Teipel, S. J., … Möller, H.-J. (2008). Structural brain alterations at different stages of schizophrenia: A voxel-based morphometric study. *Schizophrenia Research*, *104*(1–3), 44–60. doi: 10.1016/J.SCHRES.2008.06.023

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667. doi: 10.1007/s00429-010-0262-0

Modinos, G., Costafreda, S. G., van Tol, M.-J., McGuire, P. K., Aleman, A., & Allen, P. (2013). Neuroanatomy of auditory verbal hallucinations in schizophrenia: A quantitative meta-analysis of voxel-based morphometry studies. *Cortex*, *49*(4), 1046–1055. doi: 10.1016/J.CORTEX.2012.01.009

Murray, G. K., Cheng, F., Clark, L., Barnett, J. H., Blackwell, A. D., Fletcher, P. C., … Jones, P. B. (2008). Reinforcement and reversal learning in first-episode psychosis. *Schizophrenia Bulletin*, *34*(5), 848–855. doi: 10.1093/schbul/sbn078

Nakamura, M., Nestor, P. G., Levitt, J. J., Cohen, A. S., Kawashima, T., Shenton, M. E., & McCarley, R. W. (2007). Orbitofrontal volume deficit in schizophrenia and thought disorder. *Brain*, *131*(1), 180–195. doi: 10.1093/brain/awm265

Olabi, B., Ellison-Wright, I., McIntosh, A. M., Wood, S. J., Bullmore, E., & Lawrie, S. M. (2011). Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biological Psychiatry*, *70*(1), 88–96. doi: 10.1016/J.BIOPSYCH.2011.01.032

Pelayo-Terán, J. M., Pérez-Iglesias, R., Ramírez-Bonilla, M., González-Blanch, C., Martínez-García, O., Pardo-García, G., … Crespo-Facorro, B. (2008). Epidemiological factors associated with treated incidence of first-episode non-affective psychosis in Cantabria: Insights from the Clinical Programme on Early Phases of Psychosis. *Early Intervention in Psychiatry*, *2*(3), 178–187. doi: 10.1111/j.1751-7893.2008.00074.x.

Premkumar, P., Fannon, D., Sapara, A., Peters, E. R., Anilkumar, A. P., Simmons, A., … Kumari, V. (2015). Orbitofrontal cortex, emotional decision-making and response to cognitive behavioural therapy for psychosis. *Psychiatry Research: Neuroimaging*, *231*(3), 298–307. doi: 10.1016/J.PSCYCHRESNS.2015.01.013

Radewicz, K., Garey, L. J., Gentleman, S. M., & Reynolds, R. (2000). Increase in HLA-DR immunoreactive microglia in frontal and temporal cortex of chronic schizophrenics. *Journal of Neuropathology & Experimental Neurology*, *59*(2), 137–150. doi: 10.1093/jnen/59.2.137

Radua, J., Borgwardt, S., Crescini, A., Mataix-Cols, D., Meyer-Lindenberg, A., McGuire, P. K., & Fusar-Poli, P. (2012). Multimodal meta-analysis of structural and functional brain changes in first episode psychosis and the effects of antipsychotic medication. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2325–2333. doi: 10.1016/j.neubiorev.2012.07.012

Ren, W., Lui, S., Deng, W., Li, F., Li, M., Huang, X., … Gong, Q. (2013). Anatomical and functional brain abnormalities in drug-naive first-episode schizophrenia. *American Journal of Psychiatry*, *170*(11), 1308–1316. doi: 10.1176/appi.ajp.2013.12091148

Rimol, L. M., Nesvåg, R., Hagler, D. J., Bergmann, Ø, Fennema-Notestine, C., Hartberg, C. B., … Dale, A. M. (2012). Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, *71*(6), 552–560. doi: 10.1016/J.BIOPSYCH.2011.11.026

Roiz-Santiáñez, R., Pérez-Iglesias, R., Ortiz-García de la Foz, V., Tordesillas-Gutiérrez, D., Mata, I., Marco de Lucas, E., … Crespo-Facorro, B. (2011). Straight gyrus morphology in first-episode schizophrenia-spectrum patients. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *35*(1), 84–90. doi: 10.1016/J.PNPBP.2010.09.002

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., … Davatzikos, C. (2018). Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophrenia Bulletin*, *44*(5), 1035–1044. doi: 10.1093/schbul/sbx137

Salgado-Pineda, P., Baeza, I., Pérez-Gómez, M., Vendrell, P., Junqué, C., Bargalló, N., & Bernardo, M. (2003). Sustained attention impairment correlates to gray matter decreases in first episode neuroleptic-naive schizophrenic patients. *NeuroImage*, *19*(2), 365–375. doi: 10.1016/S1053-8119(03)00094-6

Sans-Sansa, B., McKenna, P. J., Canales-Rodríguez, E. J., Ortiz-Gil, J., López-Araquistain, L., Sarró, S., … Pomarol-Clotet, E. (2013). Association of formal thought disorder in schizophrenia with structural brain abnormalities in language-related cortical regions. *Schizophrenia Research*, *146*(1–3), 308–313. doi: 10.1016/J.SCHRES.2013.02.032

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience*, *10*(12), 885–892. doi: 10.1038/nrn2753

Shah, C., Zhang, W., Xiao, Y., Yao, L., Zhao, Y., Gao, X., … Lui, S. (2017). Common pattern of gray-matter abnormalities in drug-naive and medicated first-episode schizophrenia: A multimodal meta-analysis. *Psychological Medicine*, *47*(03), 401–413. doi: 10.1017/S0033291716002683

Silverstein, S. M., & Keane, B. P. (2011). Perceptual organization impairment in schizophrenia and associated brain mechanisms: Review of research from 2005 to 2010. *Schizophrenia Bulletin*, *37*(4), 690–699. doi: 10.1093/schbul/sbr052

Smieskova, R., Fusar-Poli, P., Allen, P., Bendfeldt, K., Stieglitz, R. D., Drewe, J., … Borgwardt, S. J. (2010). Neuroimaging predictors of transition to psychosis – a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *34*(8), 1207–1222. doi: 10.1016/J.NEUBIOREV.2010.01.016

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A review of reward processing and motivational impairment in schizophrenia. *Schizophrenia Bulletin*, *40*(Suppl 2), S107–S116. doi: 10.1093/schbul/sbt197

Surti, T. S., Corbera, S., Bell, M. D., & Wexler, B. E. (2011). Successful computer-based visual training specifically predicts visual memory enhancement over verbal memory improvement in schizophrenia. *Schizophrenia Research*, *132*(2–3), 131–134. doi: 10.1016/j.schres.2011.06.031

Surti, T. S., & Wexler, B. E. (2012). A pilot and feasibility study of computer-based training for visual processing deficits in schizophrenia. *Schizophrenia Research*, *142*(1–3), 248–249. doi: 10.1016/j.schres.2012.09.013

Szendi, I., Kiss, M., Racsmány, M., Boda, K., Cimmer, C., Vörös, E., … Janka, Z. (2006). Correlations between clinical symptoms, working memory functions and structural brain abnormalities in men with schizophrenia. *Psychiatry Research: Neuroimaging*, *147*(1), 47–55. doi: 10.1016/J.PSCYCHRESNS.2005.05.014

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., … Suzuki, M. (2011). Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. *PLoS ONE*. Edited by B. J. Harrison. Public Library of Science, *6*(6), 1–10. doi: 10.1371/journal.pone.0021047

Tang, J., Liao, Y., Zhou, B., Tan, C., Liu, W., Wang, D., … Chen, X. (2012). Decrease in temporal gyrus gray matter volume in first-episode, early onset schizophrenia: An MRI study. *PLoS ONE*. Edited by A. Bruce. Public Library of Science, *7*(7), e40247. doi: 10.1371/journal.pone.0040247.

Taylor, J. L., Blanton, R. E., Levitt, J. G., Caplan, R., Nobel, D., & Toga, A. W. (2005). Superior temporal gyrus differences in childhood-onset schizophrenia. *Schizophrenia Research*, *73*(2–3), 235–241. doi: 10.1016/J.SCHRES.2004.07.023

Tordesillas-Gutierrez, D., Koutsouleris, N., Roiz-Santiañez, R., Meisenzahl, E., Ayesa-Arriola, R., Marco de Lucas, E., … Crespo-Facorro, B. (2015). Grey matter volume differences in non-affective psychosis and the effects of age of onset on grey matter volumes: A voxelwise study. *Schizophrenia Research*, *164*(1–3), 74–82. doi: 10.1016/J.SCHRES.2015.01.032

van Erp, T. G. M., *et al.* (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553. doi: 10.1038/mp.2015.63

van Erp, T. G. M., *et al.* (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) consortium. *Biological Psychiatry*, *84*(9), 644–654. doi: 10.1016/J.BIOPSYCH.2018.04.023

van Erp, T. G. M., Preda, A., Nguyen, D., Faziola, L., Turner, J., Bustillo, J., … FBIRN (2014). Converting positive and negative symptom scores between

PANSS and SAPS/SANS. *Schizophrenia Research*, *152*(1), 289–294. doi: 10.1016/J.SCHRES.2013.11.013.

Venkatasubramanian, G. (2010). Neuroanatomical correlates of psychopathology in antipsychotic-naïve schizophrenia. *Indian Journal of Psychiatry*, *52*(1), 28–36. doi: 10.4103/0019-5545.58892

Vita, A., De Peri, L., Deste, G., Barlati, S., & Sacchetti, E. (2015). The effect of antipsychotic treatment on cortical gray matter changes in schizophrenia: Does the class matter? A meta-analysis and meta-regression of longitudinal magnetic resonance imaging studies. *Biological Psychiatry*, *78*(6), 403–412. doi: 10.1016/J.BIOPSYCH.2015.02.008

Vita, A., De Peri, L., Deste, G., & Sacchetti, E. (2012). Progressive loss of cortical gray matter in schizophrenia: A meta-analysis and meta-regression of longitudinal MRI studies. *Translational Psychiatry*, *2*(11), e190. doi: 10.1038/tp.2012.116

WHO (2004). *International statistical classification of diseases and related health problems*. World Health Organization.

Wylie, K. P., & Tregellas, J. R. (2010). The role of the insula in schizophrenia. *Schizophrenia Research*, *123*(2–3), 93–104. doi: 10.1016/J.SCHRES.2010.08.027

Xu, Y., Qin, W., Zhuo, C., Xu, L., Zhu, J., Liu, X., & Yu, C.(2017). Selective functional disconnection of the orbitofrontal subregions in schizophrenia. *Psychological Medicine*, *47*(09), 1637–1646. doi: 10.1017/S0033291717000101

Yassa, M., & Stark, C. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage*, *44*(2), 319–327. doi: 10.1016/j.neuroimage.2008.09.016

Neuroscience and Biobehavioral Reviews 74 (2017) 58–75

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications

Sandra Vieira[a,*], Walter H.L. Pinaya[b], Andrea Mechelli[a]

[a] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, SE5 8AF, United Kingdom
[b] Centre of Mathematics, Computation, and Cognition, Universidade Federal do ABC, Rua Arcturus, Jardim Antares, São Bernardo do Campo, SP CEP 09.606-070, Brazil

ARTICLE INFO

Article history:
Received 2 October 2016
Received in revised form
22 December 2016
Accepted 4 January 2017
Available online 10 January 2017

Keywords:
Deep learning
Machine learning
Neuroimaging
Pattern recognition
Multilayer perceptron
Autoencoders
Convolutional neural networks
Deep belief networks
Psychiatric disorders
Neurologic disorders

ABSTRACT

Deep learning (DL) is a family of machine learning methods that has gained considerable attention in the scientific community, breaking benchmark records in areas such as speech and visual recognition. DL differs from conventional machine learning methods by virtue of its ability to learn the optimal representation from the raw data through consecutive nonlinear transformations, achieving increasingly higher levels of abstraction and complexity. Given its ability to detect abstract and complex patterns, DL has been applied in neuroimaging studies of psychiatric and neurological disorders, which are characterised by subtle and diffuse alterations. Here we introduce the underlying concepts of DL and review studies that have used this approach to classify brain-based disorders. The results of these studies indicate that DL could be a powerful tool in the current search for biomarkers of psychiatric and neurologic disease. We conclude our review by discussing the main promises and challenges of using DL to elucidate brain-based disorders, as well as possible directions for future research.

Contents

* Corresponding author.
E-mail address: sandra.vieira@kcl.ac.uk (S. Vieira).

http://dx.doi.org/10.1016/j.neubiorev.2017.01.002

1. Introduction

In the last two decades, neuroimaging studies of psychiatric and neurological patients have relied on mass-univariate analytical techniques (e.g. statistical parametric mapping). These studies typically compared patients with a diagnosis of interest against disease-free individuals and reported neuroanatomical or neurofunctional differences at group level. The simplicity and interpretability of this approach have led to significant advances in our understanding of the neurobiology of psychiatric and neurological disorders. Mass-univariate analytical techniques, however, suffer from at least two significant limitations. First, statistical inferences are drawn from multiple independent comparisons (i.e. one for each voxel) based on the assumption that different brain regions act independently. This assumption, however, is not in line with our current understanding of brain function in health and disease (Fox et al., 2005; Biswal et al., 2010); for example, several psychiatric and neurological symptoms are best explained by network-level changes in structure and function rather than focal alternations (Mulders et al., 2015; Kennedy and Courchesne, 2008; Sheffield and Barch, 2016). Second, mass-univariate techniques can be used to detect differences between groups but do not allow statistical inferences at the level of the individual. In contrast, a clinician has to make diagnostic and treatment decisions about the person in front of them. These two limitations may have contributed to the limited translational impact of neuroimaging findings in everyday clinical practice so far.

In an attempt to overcome these limitations, the neuroimaging community has developed a growing interest in machine learning (ML), an area of artificial intelligence that aims to develop algorithms that discover trends and patterns in existing data and use this information to make predictions on new data. This is achieved through the use of computational statistics and mathematical optimization (Hastie et al., 2001). ML methods are multivariate and therefore take the inter-correlation between voxels into account, thereby overcoming the first limitation of mass-univariate analytical techniques. In addition, ML methods allow statistical inferences at single subject level and therefore could be used to inform diagnostic and prognostic decisions of individual patients, thereby overcoming the second limitation of mass-univariate analytical techniques (Arbabshirani et al., 2016). ML methods can be divided into two broad categories: supervised and unsupervised learning. In supervised ML, one seeks to develop a function which maps two or more sets of observations to predefined categories or values. In contrast, unsupervised methods seek to determine how the data are organized without using any a priori information supplied by the operator; here the main objective is to discover unknown structure in the data (Hastie et al., 2001).

Over the past decade, several ML methods have been applied to neuroimaging data from psychiatric and neurological patients with varying degrees of success (Arbabshirani et al., 2016; Wolfers et al., 2015). The most popular amongst these methods is Support Vector Machine (SVM), a supervised technique that works by estimating an optimal hyperplane that best separates two classes. When these classes are not linearly separable, SVM uses external functions (kernels) that map the original data into a new feature space where the data become linearly separable (Pereira et al., 2009; Vapnik, 1995). Despite its popularity, SVM has been criticised for not performing well on raw data and requiring the expert use of design techniques to extract the less redundant and more informative features (a step known as "feature selection") (LeCun et al., 2015; Plis et al., 2014). These features, rather than the original data, are then used for classification. While SVM remains a very popular technique within the neuroimaging community, an alternative family of ML methods known as deep learning (DL) (Bengio, 2009) is gaining considerable attention in the wider scientific community (Arbabshirani et al., 2016; Calhoun and Sui, 2016; LeCun et al., 2015). Deep learning methods are a type of representation-learning methods, which means that they can automatically identify the optimal representation from the raw data without requiring prior feature selection. This is achieved through the use of a hierarchical structure with different levels of complexity, which involves the application of consecutive nonlinear transformations to the raw data. These transformations result in increasingly higher levels of abstraction, where higher-level features are more invariant to the noise present in the input data than lower level ones (LeCun et al., 2015). Inspired by how the human brain processes information, the building blocks of DL neural networks – known as "artificial neurons" – are loosely modelled after biological neurons. Artificial neurons are organized in layers. A deep neural network consists of an input layer, two or more hidden layers and an output layer. The input layer comprises the data inputted into the model (e.g. voxel intensity); the hidden layers learn and store increasingly more abstract features of the data; these features are then fed to the output layer that assigns the observations to classes (e.g. controls vs. patients). Learning is achieved through an iterative process of adjustment of the interconnections between the artificial neurons within the network, much like in the human brain (Bengio, 2009). An essential aspect of DL that differentiates it from other ML methods is that the features are not manually engineered; instead, they are learned from the data, resulting in a more objective and less bias-prone process. Besides, the ability to achieve higher orders of abstraction and complexity relative to other ML methods such as SVM makes DL better suited for detecting complex, scattered and subtle patterns in the data (Plis et al., 2014).

From a historical perspective, the use of DL in scientific research can be traced back to the perceptron (i.e. the original version of the artificial neuron), which many researchers refer to as the first ML algorithm (McCulloch and Pitts, 1943). After several setbacks, the pioneering work of Warren McCulloch and Walter Pitts resulted in the development of what is now known as artificial neural networks. However, such networks were able to handle a limited number of hidden layers. It was only in the 2000s that researchers developed a new approach for training artificial neural networks that allowed the inclusion of several hidden layers resulting in greater levels of complexity (Hinton et al., 2006). This breakthrough led to the development of a new family of ML methods – known as deep learning – which has been shown to outperform previous state-of-the-art classification methods in areas such as speech recognition, computer vision and natural language processing (Krizhevsky et al., 2012; Le et al., 2012).

The use of DL could be particularly useful in the investigation of psychiatric and neurological disorders, which tend to be associated with subtle and diffuse neuroanatomical and neurofunctional abnormalities. Since high-level features can be more robust against noise in the input data, deep architectures may be more suitable to identify diagnostic and prognostic biomarkers than conventional ML methods. DL techniques might also provide an ideal tool to investigate the multi-faceted nature of psychiatric and neurological disorders since cross-modality relationships (e.g. neuroimaging and genetics) are likely to occur at an even deeper level (Plis et al., 2014). In addition to these conceptual differences, the use of DL to investigate psychiatric and neurological disorders has the practical advantage of not requiring manual feature selection (LeCun et al., 2015). Therefore, it is unsurprising that an increasing number of neuroimaging studies are using DL to elucidate the neural correlates of these disorders (e.g. Payan and Montana, 2015; Plis et al., 2014; Kim et al., 2016).

Given the insurgence of interest in DL within the field of neuroimaging, this review aims to give a brief overview of DL and potential applications to the investigation of brain-based disorders. In the first part of the review, we outline the underlying concepts of DL. To achieve this, we will use one of the simplest DL structures, i.e. the multilayer perceptron, to illustrate the steps of training and testing. This will be followed by a brief description of the most common DL architectures used in the field of neuroimaging, including stacked autoencoders, deep belief networks and convolutional neural networks. The second part of this article aims to summarise the studies that have applied DL to neuroimaging data to investigate psychiatric and neurological disorders. Finally, in the third part of the review, we discuss the main themes that have emerged from our review of the existing literature, and make a number of suggestions for future research directions.

2. Overview

Deep learning refers to the training and testing of multi-layered neural networks that are capable of learning complex structures and achieve high levels of abstraction. There are two main types of DL models which differ with respect to how the information is propagated through the network. In feedforward networks, the information is propagated through the network in just one direction, from the input to the output layer. Recurrent networks, in contrast, contain feedback connections that allow the information from past inputs to affect the current output. These connections enable the information to persist within the neural network, akin to a form of memory, and this allows the models to process sequential data, such as speech and language, in a natural way.

The implementation of DL in the context of supervised classification problems involves two main steps. In the first step, the so-called *training phase*, a subset of the available data known as the *training set* is used to optimize the network's parameters to perform the desired task (classification). In the second step, the so-called *testing phase*, the remainder subset which is known as the *test set* is used to assess whether the trained model can blind-predict the class of new observations. When the amount of available data is limited, it is also possible to run the training and testing phases several times on different training and test splits of the original data and then estimate the average performance of the model – an approach known as cross-validation. The two phases of training and testing are not a specific feature of DL but are used in conventional ML methods.

In this section, we will discuss the use of feedforward DL for classification problems. We will start with the multilayer perceptron (MLP), the simplest deep neural network (DNN) architecture, to illustrate three important aspects of deep learning – network structure, training and testing. We will then describe more complex networks, including stacked autoencoders and deep belief networks. Finally, we will describe the increasingly popular convolutional neural networks (CNN), an important adaptation of the MLP that has come to be considered the state-of-the-art for computer vision.

2.1. Multilayer perceptron

2.1.1. Network structure

MLPs are organized in a layer-wise structure where each layer stores increasingly more abstract representations of the data (Fig. 1). The first layer is the input layer where the data is entered into the model. In neuroimaging, the data can be represented as a one-dimensional vector with each value corresponding to the intensity of one voxel. The last layer is the output layer which, in the context of classification, yields the probability of a given subject belonging to one group or the other. The layers between the input and output layers are called hidden layers, with the number of hidden layers representing the depth of the network. Each layer comprises a set of artificial neurons or "nodes" (Fig. 1a) in which each neuron is fully connected to all neurons in the previous layer (Fig. 1b). Each connection is associated with a weight value, which reflects the strength and direction (excitatory or inhibitory) of each neuron input, much like a synapse between two biological neurons.

Unlike SVM, which relies on expert designed transformations to handle nonlinearly separable classes, the structure of neural networks itself allows the transformation of the input space. The consecutive layers perform a cascade of nonlinear transformations that distort the input space allowing the data to become more easily separable (Fig. 2). The optimal number of layers and nodes within each layer are not estimated as part of the learning process itself but are defined *a priori*. These *a priori* parameters, which are not optimized during the training, are called hyperparameters. It should be noted that the development of algorithms to find optimum values of these hyperparameters is an active area of research, and that at present there are no fixed rules (Bergstra et al., 2011; Gelbart et al., 2014).

2.1.2. Training

Traditionally, neural networks can learn through a gradient descent-based algorithm. The gradient descent algorithm aims to find the values of the network weights that best minimise the error (difference) between the estimated and true outputs. Since MLPs can have several layers, in order to adjust all the weights along the hidden layers, it is necessary to propagate this error backward (from the output to the input layer). This propagation procedure is called backpropagation, and allows the network to estimate how much the weights from the lower layers need to be changed by the gradient descent algorithm. Initially, when a neural network is trained,

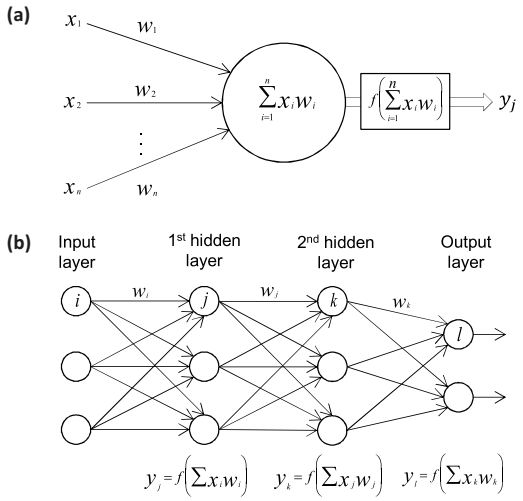


Fig. 1. (a) The building block of deep neural networks – artificial neuron or node. Each input x_i has an associated weight w_i. The sum of all weighted inputs, $\Sigma x_i w_i$, is then passed through a nonlinear activation function f, to transform the pre-activation level of the neuron to an output y_j. For simplicity, the bias terms have been omitted. The output y_j then serves as input to a node in the next layer. Several activation functions are available, which differ with respect to how they map a pre-activation level to an output value. The most commonly activation functions used are the rectifier function (where neurons that use it are called rectified linear unit (ReLU)), the hyperbolic tangent function, the sigmoid function and the softmax function. The latter is commonly used in the output layer as it can compute the probability of multiclass labels. (b) Example of a feedforward multilayer neural network (also referred to as multilayer perceptron) with two classes, in which the nodes in one layer are connected to all neurons in the next layer (fully connected network). For each neuron j in the first hidden layer, a nonlinear function is applied to the weighted sum of the inputs. The result of this transformation (y_j) serves as input for the second hidden layer. The information is propagated through the network up to the output layer, where the softmax function yields the probability of a given observation belonging to each class.

the weights are set at random. When the training set is presented to the network, this forward propagates the data through the nonlinear transformation along the layers. The estimated output is then compared to the true output, and the error is propagated from the output towards the input, allowing the gradient descent algorithm to adjust the weights as required. The process continues iteratively until the error has reached its minimum value. The backpropagation algorithm does not work well with the original models of DNNs that were based on sigmoid and hyperbolic tangent nonlinearities. In these models, the information of the error becomes increasingly smaller as it propagates backward from the output to the input layer, to a point where initial layers do not get useful feedback on how to adjust their weights – an issue known as the vanishing gradient problem. Therefore, initially, the use of backpropagation yielded poor solutions for networks with three or more hidden layers (Schmidhuber, 2015). In 2006, however, Hinton and colleagues put forward the idea of "greedy layerwise training", which consists of two steps: 1) an unsupervised step, where each layer is trained individually and 2) a supervised step, where the previously trained layers are stacked, one additional layer is added to perform the classification (the output layer), and the whole network parameters are fine-tuned (Hinton et al., 2006). This breakthrough led to the fast-growing interest in deep learning and enabled the development of at least two types of pre-trained networks that have shown promising results: stacked autoencoders and deep belief networks. It should be noted that these methods are not actual classifiers themselves; instead, they are networks that are pre-trained to learn useful patterns in the data and then fed to a real classifier at the final layer. These two types of networks and their unique characteristics are described in Section 2.2 and 2.3.

2.1.3. Testing

The performance of a deep neural network can be evaluated by several performance measures, such as sensitivity, specificity, accuracy and F-score. Sensitivity refers to the proportion of true positives correctly identified (e.g. the proportion of subjects that were predicted as patient and are true patients), and specificity refers to true negatives correctly identified (e.g. the proportion of subjects that were predicted as healthy controls and are true healthy controls). The accuracy of a classifier represents the overall proportion of correct classifications. The statistical significance of this overall accuracy can be tested using parametric tests such as permutation testing, which measures how likely the observed accuracy would be obtained by chance. Metrics such as F-score and balanced accuracy, which take into account each group's sample size, are particularly useful in cases where classes are unbalanced. The F-score is a measure that combines precision or positive predictive value (proportion of individuals classified as cases were actually cases) and sensitivity (proportion of true cases correctly classified as such). Balanced accuracy, on the other hand, corresponds to the average accuracy obtained on either class (Brodersen et al., 2010).

2.1.4. Risk of overfitting and possible strategies

Due to the use of multiple nonlinear transformations, deep networks are highly complex models that involve the estimation of a very large number of parameters. This can lead to the model learning particular fluctuations in the training data that are irrelevant

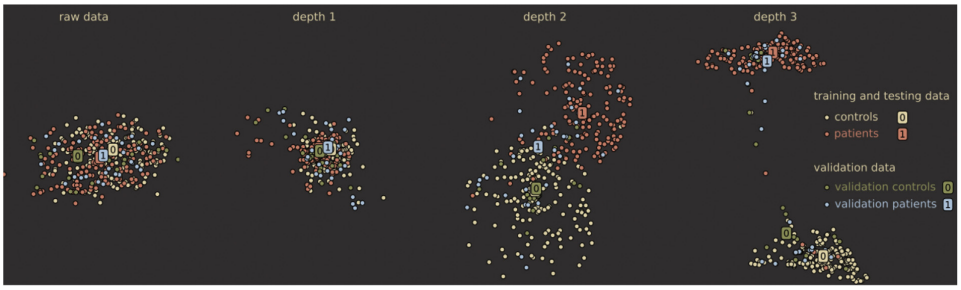


Fig. 2. Effect of the depth of the model. Each dot corresponds to a neuroimage-based data visualized in a two-dimensional map. With more hidden layers, the data becomes more easily separable due to nonlinear transformations along the network (Plis et al., 2014).

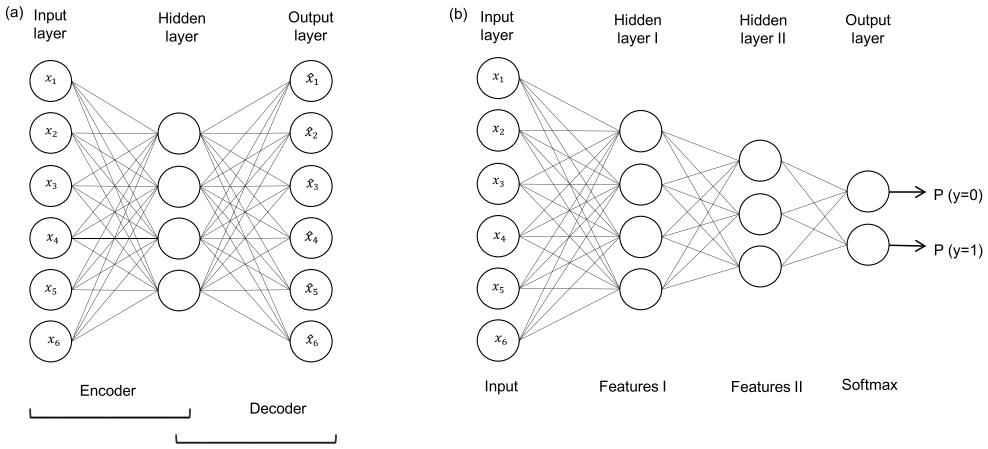


Fig. 3. (a) Shallow or simple autoencoder. In its shallow structure, an autoencoder is comprised of an input layer, that represents the original data (e.g., pixels in an image), one hidden layer that represents the transformed data, and an output layer that reconstructs the original input data. (b) Stacked autoencoder. Two simple autoencoders are stacked with a 2-class softmax classifier as the final layer. From each simple autoencoder, the output layer is discarded, and the hidden layer is used as the input layer for next autoencoder.

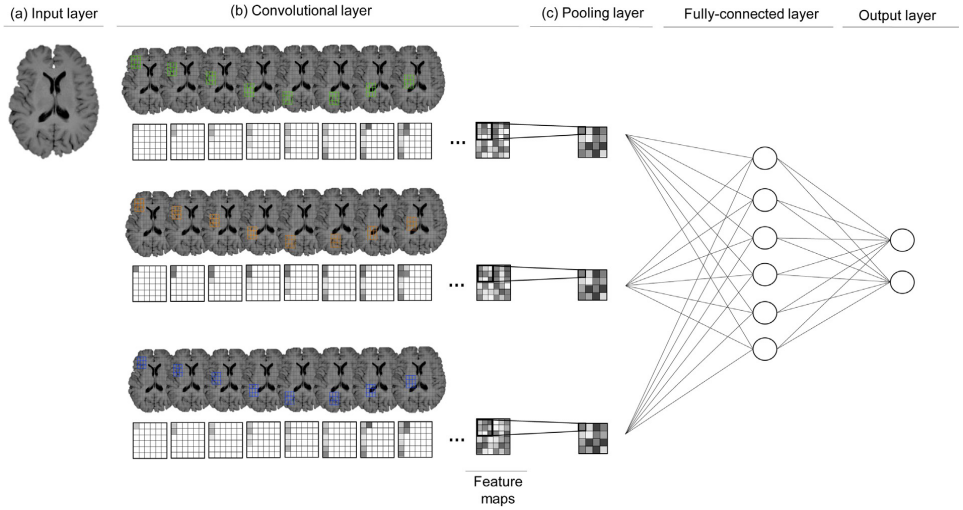


Fig. 4. Generic structure of a CNN. For illustrative purpose, this example only has one layer of each type; a real-world CNN, however, would have several convolutional and pooling layers (usually interpolated) and one fully-connected layer. (a) Input layer. In its simplest way, the data is inputted into the network in such a way that each pixel corresponds to one node in the input layer. (b) Convolutional layer. A 3 × 3 filter or kernel (in green) is used to multiply the spatially corresponding 3 × 3 nodes in the image. The resulting weighted sum is then passed through a nonlinear function to derive the output value of one node in the feature map. The repetition of this same operation across all possible receptive fields results in one complete feature map. The same procedure with different kernels (in orange and blue) will result in separate complete feature maps. (c) Pooling layer. The size of each feature map can be reduced by taking the maximum value (or average) from a receptive field in the previous layer. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for the purpose of classification – an issue known as "overfitting". When this happens, the model will perform very well on the training data but will not be able to replicate its performance on unseen data (Srivastava et al., 2014). The risk of overfitting is particularly high in the context of neuroimaging, where the number of data points (e.g. number of voxels) for a subject is much larger than the total number of subjects, resulting in high-dimensional data (Arbabshirani et al., 2016). However, there are a number of strategies that can be used to minimise the risk of overfitting, collectively known as "regularization". A first strategy involves the use of weight decays (e.g., L1 and L2 norms) to penalise models with very high weights. It has been observed that extreme (very low or very high) weight values in a ML model are symptomatic of the model trying to learn the regularities of the data perfectly (Moody et al., 1995). By forcing weights to remain low, the network becomes less dependent on the training data and is able to better generalise to unseen data (Nowlan and Hinton, 1992). A second strategy, known as dropout, consists of temporarily removing

a random number of nodes and their respective incoming and outgoing connections from the network during training. This means that the contribution of dropped-out neurons to the activation of downstream neurons is temporally removed on the forward pass and that any weight updates are not applied to these neurons on the backward pass. The aim of dropout is to extract different sets of features that can independently produce a useful output, thereby allowing higher levels of generalizability (Srivastava et al., 2014).

2.2. Autoencoders

Autoencoders are a special case of feedforward networks which comprise of two main components. The first component, i.e. the "encoder", learns to generate a latent representation of the input data, whereas the second component, i.e. the "decoder", learns to use these learned latent representations to reconstruct the input data as close as possible to the original (Fig. 3a) (Vincent et al., 2010).

Since an autoencoder does not make use of labels, its training is an unsupervised learning process. In its shallow structure, an autoencoder is comprised of three layers: an input layer, one hidden layer and an output layer. The training to perform the input-copying task can be useful to extract meaningful features of the input data. This automatic feature extraction can be performed using an error function (or loss function) that encourages the model encoder to have specific characteristics, such as sparsity of the representation (sparse autoencoders) and robustness to noise (denoising autoencoders). Since autoencoders are automatic features extractors, they can also be stacked to create a deep structure to increase the level of abstraction of learned features. In this case, the network is pre-trained, i.e. each layer is treated as a shallow autoencoder, generating latent representations of the input data. These latent representations are then used as input for the subsequent layers before the full network is fine-tuned using standard supervised learning (Fig. 3b) (Larochelle et al., 2007).

2.3. Deep belief networks

Deep belief networks (DBNs), proposed by Hinton et al. (2006), are technically the first DL models. Similar to stacked autoencoders, DBNs are comprised of stacked shallow feature extractors, known as restricted Boltzmann machines (RBMs). An RBM is composed by only two layers: a visible layer and a hidden layer. Just like autoencoders, RBMs also aim to learn and extract useful features from the data. However, RBMs differ from autoencoders with regards to their training processes. RBMs can be interpreted as a stochastic neural network. Therefore, instead of using deterministic functions and the reconstruction error (like the autoencoders), the RBM uses the maximum-likelihood estimation to find a stochastic representation of the input in its hidden layer (latent features). To do this, RBMs are usually trained using a gradient descent algorithm, with the likelihood gradient being performed by an approximation algorithm known as contrastive divergence (Hinton et al., 2006). Here the input data, stored in the visible layer, are propagated to the hidden layer as in a feedforward network, and the resulting sum of the weighted inputs provides a measure of the neuron activation probability. The activation of hidden neurons can be thought of as the network's internal representation of the data, which is then propagated back to the visible layer in an attempt to reconstruct the input data from the network's internal representation. The network, therefore, learns by adjusting the weights based on the discrepancy between the true and reconstructed data. Similarly to autoencoders, RBMs can be stacked to create a deep network, where the hidden layer representation of one RBM serves as input layer for the following RBM, and the network can learn higher-level features from lower-level ones to arrive at an abstract representation of the data. Furthermore, the neural network corresponding to a trained DBN can be augmented by adding an output layer, where units represent the labels corresponding to the input sample. This results in a standard neural network for classification that can be further trained using supervised learning algorithms.

2.4. Convolutional neural networks

Convolutional neural networks (CNNs) are a special type of feedforward neural networks that were initially designed to process images, and as such are biologically-inspired by the visual cortex (LeCun et al., 1998). In addition to the input and output layers, CNN can comprise of three types of layers: a convolutional layer, a pooling layer, and a fully-connected layer (Fig. 4).

The convolutional layer is organized in several feature maps. Every neuron in a feature map is connected to a fixed set of neurons in a local region of the previous layer – the *receptive field* – in such a way that the whole image is covered ("local connectivity"). Within the same feature map, the connections between each neuron and the corresponding *receptive field* share the same weights, whereas different feature maps use different sets of weights ("weight sharing"). As a result of this architecture, a feature map can be thought of as a "feature detector" that scans the whole image for the same pattern. This pattern is usually known as the kernel. Kernels in a CNN are learned during the training process, as opposed to in SVM, where they are defined a priori. In a network with several convolutional layers, each layer codes for increasingly more abstract features (e.g. lines → edges → eyes → face). The pooling layer simply reduces the number of neurons of the previous convolutional layer. The fully-connected layers are similar to the hidden layers from the conventional MLP where the neurons are connected to all neurons from the previous layer. All combined, the properties of CNN (local connectivity, weight sharing and pooling) result in a significant reduction in the number of parameters, which in turn decreases the likelihood of overfitting, and alleviates computational processing.

3. Review of DL studies of psychiatric or neurological disorders

In order to identify previous applications of DL in neuroimaging studies of psychiatric or neurological disorders, a search was conducted on 1st August 2016 across several databases (PubMed, IEEE Xplore, Scopus and ArXiv) using the following search terms: ("deep learning" OR "deep architecture" OR "artificial neural network" OR "autoencoder" OR "convolutional neural network" OR "deep belief network") AND (neurology OR neurological OR psychiatry OR psychiatric OR diagnosis OR prediction OR prognosis OR outcome) AND (neuroimaging OR MRI OR "Magnetic Resonance Imaging" OR "fMRI" OR "functional Magnetic Resonance Imaging" OR PET OR "Positron emission tomography"). This review did not include EEG studies, although there is some evidence that DL can also be used with this type of data, particularly in epilepsy (Page et al., 2014). The initial search yielded a total of 172 articles. As the next step, we screened and cross-referenced these articles for studies that had applied a deep learning model to neuroimaging data to investigate a psychiatric or neurologic condition; this identified a total of 25 articles which were relevant to our review. We organized these articles as follows: i) *diagnostic studies*, which aimed to classify patients from healthy controls, ii) *studies on conversion to illness*, which used baseline scans from individuals identified as being at high risk of developing a psychiatric or neurologic disorder to predict subsequent transition to the illness, and finally iii) *studies predicting treatment response,* which used baseline scans from individuals with a neurological or psychiatric diagnosis to predict

Table 1
Diagnostic studies.

Authors, year	Sample size	Technique	Features	Previous feature selection	DL architecture	Comparison	Accurac
Gupta et al. (2013)[a]	AD = 200 MCI = 411 HC = 232	sMRI	WB voxel-level	No	Sparse AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	94.7 86.4 88.1 85.0
Payan and Montana (2015)[a]	HC = 755 AD = 755 MCI = 755	sMRI	WB voxel-level	No	Sparse AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	95.4 92.1 86.8 89.5
Hosseini-Asl et al. (2016) [a,b]	HC = 70* AD = 70* MCI = 70*	sMRI	WB voxel-level	No	AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	97.6 90.8 95.0 89.1
Chen et al. (2015)[a]	HC = 123 AD = 94 MCI = 121	sMRI	WB voxel-level	Yes	SAE	HC vs. AD HC vs. MCI	89.0 81.7
Liu et al. (2015a)[a]	HC = 204 AD = 180 MCI = 374	sMRI	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	82.6 72.0
Gao and Hui (2016)	HC = 117 AD = 51 Lesions = 118	CT	WB voxel-level	No	CNN	HC vs. AD vs. Lesion	87.7
Sarraf and Tofighi (2016)[a]	HC = 15 AD = 28	rsfMRI	WB voxel-level	No	CNN	HC vs. AD	96.9
Suk et al. (2016)[a]	HC = 31 MCI = 31	rsfMRI	WB region-level	Yes	DAE	HC vs. MCI	72.6
	HC = 25 MCI = 12	rsfMRI	WB region-level	Yes	DAE	HC vs. MCI	81.1
Hu et al. (2016)[a]	HC = 52 MCI = 48	rsfMRI	WB region-level	No	SAE	HC vs. MCI	87.5
Han et al. (2015)[a]	HC = nr AD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. AD	80.0
Liu et al. (2015a)[a]	HC = 77 AD = 85 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	91.4 82.1
Suk et al. (2014)[a]	HC = 101 AD = 93 MCI = 204	sMRI & PET	WB region-level	Yes	DBM	HC vs. AD HC vs. MCI	94.9 80.6
Liu et al. (2014)[a]	HC = 77 AD = 65 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	87.8 76.9
Suk et al. (2015b)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	HC vs. AD HC vs. MCI HC vs. AD vs. MCI	95.1 80.1 62.9
	HC = 229 AD = 198 MCI = 403	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	HC vs. AD HC vs. MCI HC vs. AD vs. MCI	90.3 70.9 57.7
Liu et al. (2015b)[a]	HC = 77 AD = 85 MCI = 169	sMRI & PET & MMSE	WB region-level	Yes	SAE	HC vs. AD HC vs. AD vs. MCI	90.1 59.2
Suk et al. (2015a)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI AD vs. MCI	98.8 90.7 83.7
Li et al. (2014)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	HC vs. AD HC vs. MCI	91.4 77.4
Suk and Shen (2013)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	HC vs. AD HC vs. MCI	95.9 85.0
Han et al. (2015)[c]	HC = nr ADHD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. ADHD	65.0
Deshpande et al. (2015)[c]	HC = 744 ADHD-C = 260 ADHD-I = 173	rsfMRI	WB region-level	Yes	FCC	HC vs. ADHD-C HC vs. ADHD-I ADHD-C vs. ADHD-I	~90.0 ~90.0 95.0

Table 1 (*Continued*)

Authors, year	Sample size	Technique	Features	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Kuang et al. (2014)[c]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC) ROI (VC) ROI (CC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	37.4 to 71.8[***] 34.4 to 68.8[***] 37.1 to 72.7[***]
Kuang and He (2014)[c]	HC = 42 to 95 ADHD-C = 0 to 77 ADHD-I = 0 to 44 ADHD-H = 0 to 6	rsfMRI	ROI (PFC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	44.4 to 80.9[***]
Hao et al. (2015)[c]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC, VC, SSC and CC combined)	Yes	DBaN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	48.9 to 72.7[***]
Plis et al. (2014)	HC = 191 SZ and FEP = 198	sMRI	WB voxel-level	No	DBN	HC vs. SZ	91[**]
Kim et al. (2016)[d]	HC = 50 SZ = 50	rsfMRI	WB region-level	Yes	SAE	HC vs. SZ	85.8
Munsell et al. (2015)	HC = 48 TLE = 70	DTI	WB region-level	No	SAE	HC vs. TLE	69.0
Yang et al. (2014)	HC = 31 SCA2 = 4 SCA6 = 27 AT = 18	sMRI	ROI (Cerebellum)	No	SAE	HC vs. SCA2 vs. SCA6 vs. AT	86.3

[a] ADNI dataset.
[b] CADDementia dataset.
[c] ADHD-200 dataset.
[d] COBRE dataset.
[*] Sample sizes for the fine-tuning stage only (pre-training included an additional 386 samples).
[**] F-score.
[***] Range of accuracies obtain from the different datasets used; HC, healthy controls; SZ, schizophrenia, FEP, first episode psychosis; ADHD, attention deficit/hyperactive disorder; ADHD-C, attention-deficit/hyperactive disorder combine subtype; ADHD-I, attention-deficit/hyperactive disorder inattentive subtype; ADHD-H, attention-deficit/hyperactive disorder hyperactive subtype; SCA2, spinocerebellar ataxia type 2; SCA6, spinocerebellar ataxia type 6; AT, ataxia-telangiectasia; TLE, temporal lobe epilepsy; AD, Alzheimer's disease; MCI, mild cognitive impairment; CC, cingulate cortex; VC, visual cortex, PFC, pre-frontal cortex; SSC, somatosensory cortex; sMRI, structural MRI; rsfMRI, resting-state functional MRI; CT, computed tomography; PET, Positron emission tomography; DTI, diffusion tensor imaging; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; AE, autoencoder, SAE, stacked autoencoder; FCC, fully-connected cascade; DBN, deep belief network, DBaN, deep Bayesian network; CNN, convolutional neural network; DAE, deep autoencoder; DBM, deep Boltzman machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron; nr, not reported.

subsequent treatment response. These studies are summarised in Tables 1, 2 and 3 which provide the following information: sample size; type of data used as input; whether a whole brain (WB) or region of interest (ROI) approach was used; whether the information inputted into the model comprised of voxel or region-level features; whether feature selection was or was not used before inputting the data into the model; general type of DL architecture; diagnostic groups being investigated; and accuracy. Whenever performed, we also report the accuracies obtained for multiclass classifications, which involve discriminating between more than two classes (e.g. healthy controls vs. mild cognitive impairment vs. Alzheimer's disease).

3.1. Diagnostic studies

Studies using DL to classify psychiatric or neurological patients from healthy individuals have used a range of neuroimaging modalities including structural MRI (sMRI), resting-state fMRI (rsfMRI), positron emission tomography (PET) and a combination of different modalities (multimodal studies) (see Table 1). From Table 1 it can be seen that the vast majority of these studies were carried out in Alzheimer's disease (AD) and its prodromal stage, mild cognitive impairment (MCI). In addition, a smaller number of studies examined psychosis, attention deficit/hyperactivity disorder (ADHD), cerebellar ataxia and temporal lobe epilepsy (TLE). Within each diagnostic category, we first give an overview of the studies that have used a single neuroimaging modality, followed by studies that employed a multimodal approach and, finally, studies that have combined neuroimaging and clinical data within a single classifier.

3.1.1. Mild Cognitive Impairment and Alzheimer Dementia

In one of the first studies using DL in AD and MCI, Gupta et al. (2013) argued that, since (i) natural images and brain imaging have similar, and therefore interchangeable, low-level features (e.g. lines and corners) and (ii) natural images, contrary to neuroimaging, are abundant, then natural images could be used to learn low level features which could then be used to identify lesions along the surface and ventricles of the brain. This process, whereby the features

Table 2
Conversion to illness.

Authors, year	Sample size	Technique	WB voxel-level/WB region-level/ROI	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Liu et al. (2015a)[a]	HC = 204 AD = 180 MCI-C = 160 MCI-NC = 214	sMRI	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	46.3
Suk et al. (2014)[a]	MCI-C = 76 MCI-NC = 128	sMRI & PET	WB region-level	Yes	DBM	MCI-NC vs MCI-C	71.6
Liu et al. (2015a)[a]	HC = 77 AD = 85 MCI-C = 67 MCI-NC = 102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	53.8
Liu et al. (2014)[a]	HC = 77 AD = 65 MCI-C = 67 MCI-NC = 102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	47.4
Suk et al. (2015b)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C	74.2
						AD vs MCI-C vs MCI-NC vs HC	53.7
	AD = 51 HC = 52 MCI-C = 167 MCI-NC = 236	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C	73.9
						AD vs MCI-C vs MCI-NC vs HC	47.8
Li et al. (2014)[a]	HC = 52 AD = 198 MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	MCI-NC vs MCI-C	57.4
Suk and Shen (2013)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	MCI-NC vs MCI-C	75.8
Suk et al. (2015a)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	MCI-NC vs MCI-C	83.3

[a] ADNI dataset; HC, healthy controls; AD, Alzheimer's disease; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; sMRI, structural MRI; PET, Positron Emission Tomography; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; SAE, stacked autoencoder; DBM, deep Boltzmann machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron.

Table 3
Treatment outcome.

Authors, year	Sample size	Technique	WB voxel-level/WB region-level/ROI	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Munsell et al. (2015)	TLEns = 41 TLEs = 29	DTI	WB region-level	No	SAE	TLEns vs TLEs	57.0

HC, healthy controls; TLE-ns, temporal lobe epilepsy without seizures; TLE-ns, temporal lobe epilepsy with seizures; DTI, diffusion tensor imaging.

learned in one set of data are used to solve a problem in another set of data, is known as "transfer learning". Based on this premise, the authors pre-trained a sparse autoencoder to learn features from natural images, which were then applied to structural MRI data via a CNN, achieving a classification accuracy of 94.7% for AD versus controls, 86.4% for MCI versus controls and 88.1% for AD versus MCI. Consistent with the authors' hypothesis, this method outperformed the one where the learned features were extracted from the neuroimaging data (93.8%, 83.3% and 86.3% for the same comparisons, respectively). However, a few years later and using a similar approach, Payan and Montana (2015) found comparable classification accuracies using features that were learned from the structural MRI data itself. This could potentially be explained by the fact that Payan and Montana (2015) used a much larger sample, as well as by the fact that authors used 3D brain images, as opposed to 2D, which possibly contain more useful patterns for classification. Indeed, Payan and Montana (2015) reported that, in general, the models based on 3D outperformed those based on 2D brain images (AD vs. HC (2D/3D) = 95.4%/95.4%; AD vs. MCI (2D/3D) = 82.2%/86.8%; MCI vs. HC (2D/3D) = 90.1%/92.1%). The best accuracy (97.6%) from single modality studies came from Hosseini-Asl et al. (2016), who also used transfer learning. Instead of extracting features from natural images and then fine-tuning the model on Alzheimer's patients and controls, as seen in Gupta et al. (2013); Hosseini-Asl et al. (2016) used one Alzheimer's dataset for pre-training and another independent Alzheimer's dataset to fine-tune the model. By performing the pre-training on an Alzheimer's dataset, this approach allowed for the network to extract generic features related to AD biomarkers, such as the ventricular size, hippocampus shape, and cortical thickness as opposed to more generic low-level features as in Gupta et al. (2013). By using two independent samples during the complete learning process, the final learned features for classification

are much less dataset-specific, and should therefore be more generalizable. The final model's architecture was also deeper than in previous studies, which probably also contributed to the high accuracy. Taken collectively, these studies suggest that the application of DL to structural MRI data allows the classification of individuals with AD and MCI with high levels of accuracy. Consistent with the increasing popularity of CNN models, studies that have applied either CNN or a combination of AE and CNN have shown better performances compared to those using only AE, although it should be noted that the former group of studies tended to have larger samples than the latter group. In addition, and similar to the trend reported in computer vision competitions and research, the best performances were obtained by the deepest CNN models.

Studies of AD and MCI using resting-state imaging have also achieved promising results. For example, Han et al. (2015) designed a hierarchical convolutional sparse autoencoder (HCSAE), which essentially extracts the most discriminating features from the resting-state data and encodes them in a convolutional manner. This particular arrangement allows for the extraction of the most useful information while conserving abundant detail. The final model classified AD and controls with an 80.0% accuracy and significantly outperformed SVM, which only yielded an accuracy of 50% (Fig. 4). While this is a promising result, the model assumed that functional networks were statistic over time – an assumption which underlies the vast majority of ML applications to resting-state neuroimaging data. However, recent studies have shown that the network-level functional organization of the brain is dynamic rather than static (Hutchison et al., 2013). Suk et al. (2016) have addressed this issue by developing an approach which classifies people with MCI and healthy controls using a deep autoencoder to extract hierarchical nonlinear relations among brain regions, whilst modelling the inherent functional dynamics of resting-state data. This was also one of the few studies in which the same DL model was tested against and surpassed other competing models in two independent datasets (72.6% for dataset 1 and 80.0% for dataset 2), thus providing evidence of replicability, a crucial feature for diagnostic tools. In line with the studies using structural imaging, the best performance for the classification of AD patients with resting-state data was also obtained by a CNN model with an accuracy of 96.9% (Sarraf and Tofighi, 2016). These studies provide initial evidence that brain activity at resting state can be useful in identifying MCI and AD patients. We note that, compared to the performances obtained from structural data, DL models applied to functional data seem to perform worse. This discrepancy could be explained by the substantial difference in sample size between the two types of studies – while the *smallest* study using structural data included 140 subjects (Hosseini-Asl et al., 2016) the *largest* study using functional data included 62 subjects (Suk et al., 2016).

With regards to multimodal studies, Liu et al. (2014) applied a stacked autoencoder (SAE) to structural and PET data and successfully distinguished AD and MCI from controls with an accuracy of 87.8% and 76.9%, respectively. Using a very similar dataset, the same team (Liu et al., 2015a) achieved a better performance by designing a model where the hidden layers were able to infer the correlations between sMRI and PET, thus better capturing the synergy between the two modalities. This model classified AD and MCI against controls with an accuracy of 91.4% and 82.1%, respectively. Interestingly, the application of the same model to a structural data alone resulted in less impressive accuracies of 82.6% and 72% for AD and MCI, respectively. This discrepancy suggests that the integration of structural and functional data may improve classification accuracy. However, this conclusion should be drawn with great caution since that the authors did not report classification accuracy for PET data alone.

Finally, four studies have tried combining neuroimaging data with clinical information to build a more robust classification model. For example, Suk and Shen (2013) used a SAE to extract latent features from neuroimaging data (sMRI, PET and CSF), which were then used to predict clinical data (measured using the Mini-Mental State Examination – MMSE – and Alzheimer's Disease Assessment Scale's cognitive subscale – ADAS-cog) and class labels. As the final step, the resulting learned features were used to classify AD and MCI from healthy individuals with an accuracy of 95.9% and 85.0%, respectively. Notably, two more studies (Li et al., 2014; Suk et al., 2015a) that have used the same exact sample (taken from the publicly available dataset ADNI; Alzheimer's Disease Neuroimaging Initiative) and the same types of data (sMRI, PET, CSF, MMSE and ADAS-cog) have also reported high accuracies for both AD and MCI despite using different implementations of DL. In general, studies combining clinical with neuroimaging data have, in general, reported higher accuracies than studies using single modality or multiple neuroimaging modalities. This is in line with previous studies using conventional ML methods (e.g. Willette et al., 2014; Moradi et al., 2015; Zhang and Shen, 2012) and highlights the usefulness of adding clinical information in the classification of AD and its prodromal phase.

3.1.2. Attention-deficit/hyperactive disorder

With regards to attention-deficit/hyperactivity disorder (ADHD), all five studies included here have used resting-state neuroimaging data. For example, Deshpande et al. (2015) applied a fully connected cascade artificial neural network – a variation of the multilayer perceptron – to functional connectivity from ADHD and healthy controls. The model successfully distinguished between the inattentive and combined subtypes from healthy controls with an accuracy of 90% for both comparisons, while the two subtypes were discriminated with an accuracy of 95%. Connections between frontal areas and the cerebellum were identified as the most discriminating features. There is also evidence that healthy children and children diagnosed with three different ADHD subtypes (inattentive, hyperactive and combined) can be distinguished in one single model using a multiclass approach, without the need to perform binary classifications between healthy controls and each ADHD subtypes. This evidence comes from three studies that have used data from different sites taken from the ADHD-200 consortium, a data-sharing platform aimed at understanding the neural basis of ADHD (Milham et al., 2012). Kuang et al. (2014) attempted to discriminate between healthy controls and ADHD subtypes (inattentive, hyperactive and combined) using data acquired from three different sites. Rather than looking at the whole brain, the authors first parcellated the brain and trained different DBNs for each brain area to examine which part of the brain best discriminated ADHD (regardless of subtypes) from healthy controls. A 4-way DBN was then performed for the each best discriminating area – prefrontal (PFC), cingulate (CC) and visual (VC) cortex – in each one of the three datasets separately (dataset 1: PFC = 37.4%, CC = 37.1%, VC = 34.4%; dataset 2: PFC = 54.0%, CC = 54.0%, VC = 51.2%; dataset 3: PFC = 71.8%, CC = 72.7%, VC = 68.8%). Kuang and He (2014) partially replicated these findings by applying the same DL approach to functional measures of the prefrontal cortex; this allowed a 4-way classification accuracy of 44.4%, 55.6% and 80.9% in three independent samples from the ADHD-200 consortium. Finally, Hao et al. (2015) identified the most discriminating areas – prefrontal, cingulate, somatosensory and visual cortex – and then combined them within a single model. The resulting input data were put through a deep Bayesian network (DBaN), where a DBN was used to reduce the dimensionality of the data and a Bayesian network was used to extract the relationships between the data. The resulting model achieved a 4-way classification accuracy of 48.8%, 54.0% and 72.7% for three independent samples also taken from the ADHD-200 consortium. These three studies suggest that DL can be used to

solve multiclass classifications problems, as all performances were well above chance level (25% for a classification with 4 classes). In addition, these studies suggest that DL can extract meaningful information from patterns of brain functioning to classify ADHD from controls and, more notably, to differentiate between ADHD subtypes. Nevertheless, we note that all four studies conducted in ADHD had unbalanced sample sizes between classes. For example, in Kuang et al. (2014), there were just between 2 and 5 children in the Inattentive subtype within each site, while the number of healthy children ranged from 69 to 110 per site. Similarly, each site in Kuang and He (2014) did not include any participants on at least one ADHD subtype which may have introduced a bias in the 4-way classification performed across all sites. With the exception of Hao et al. (2015) which reported sensitivity and specificity, all studies assessed model performance by estimating the overall accuracy. This metric is simply the proportion of participants correctly identified, and therefore does not take the unbalance between classes into account; this means that it is possible to have a good overall accuracy even if several participants from a class are misclassified (or even if all participants from a class are misclassified if the sample size for that class is very small compared to the total sample size). Therefore, given the highly imbalanced sample sizes, the possibility that the performances reported in these studies are inflated cannot be ruled out. This possibility is supported by the observation of much lower sensitivities (43.9%, 22.9% and 55.6% for each site) than specificities (68.8%, 87.7% and 83.0%), in Hao et al. (2015).

3.1.3. Psychosis

With respect to psychosis, two studies have been performed with promising results. Using structural MRI data from four independent studies, Plis et al. (2014) applied a DBN to the original pre-processed images obtaining an impressive F-score of 91%. While this was a highly promising result, the patients group included both first episode and chronic schizophrenia patients, which could have diluted the models' performance. More recently, Kim et al. (2016) extracted functional connectivity patterns obtained from resting-state functional MRI of individuals diagnosed with schizophrenia and healthy controls and performed a series of experiments with an SAE-based model, in which different hyperparameters were tested. The proposed model consisted of an SAE with weight sparsity control, i.e. only a random selection of neurons in a given layer was activated, that classified schizophrenia patients and controls with an accuracy of 85.5%, outperforming SVM by a margin of 8.1%. Consistent with the literature on brain functional abnormalities in schizophrenia (Kühn and Jürgen, 2013; van der Meer et al., 2010), the most relevant features for the classification were the functional connectivity between the thalamus and the cerebellum, the frontal and temporal areas and between the precuneus/posterior cingulate cortex and the striatum. Despite this encouraging result, the sample sizes for each class were modest (50 for each group) and, therefore, it is not clear how well these findings will generalise to a different sample. Nevertheless, both studies suggest that DL can effectively classify psychosis patients on the basis of neuroanatomical and neurofunctional information. Despite the evidence that structural and functional data provide complementary information on the neural basis of psychosis (Cabral et al., 2016; Radua et al., 2012; Schultz et al., 2012), to date there have been no DL studies using a multimodal approach in psychosis. In addition, despite the evidence that psychosis, similar to AD, is preceded by a prodromal stage (Yung et al., 2005), there have been no studies applying DL to neuroimaging data to classify individuals at high risk of developing psychosis from healthy controls or distinguishing between high risk individuals who will and will not develop the illness.

3.1.4. Temporal lobe epilepsy

One study examined the potential of DL to classify healthy individuals and patients diagnosed with temporal lobe epilepsy (TLE) from diffusion-weighted images (DWI) (Munsell et al., 2015). A stacked autoencoder was used to extract meaningful features from patients' connectome while SVM was chosen as the classifier. Deep learning was suggested as an attractive ML alternative because it is capable of encoding latent, nonlinear relationships in high dimension data. This combination yielded a relatively modest accuracy of 69%. In addition, this model was outperformed by another approach where features were extracted using a well-known linear automated method (ElasticNet) instead, which achieved an accuracy of 80%. This discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. Given the high complexity resulting from the numerous parameters to be estimated, DL models are more prone to overfitting (high performance on the training data while performing poorly on unseen data) than conventional ML approaches. One standard solution, that the authors did not use, is to address this issue is by tuning the level of model complexity and penalizing highly intricate ones in order to have better generalizing models.

3.1.5. Cerebellar ataxia

One study was conducted in cerebellar ataxia (CA), a neurodegenerative disorder that affects mainly the cerebellum, with multiple genetics variations each with its characteristic pattern of anatomical degeneration. Yang et al. (2014) applied a stacked AE to T1-weighted images of the cerebellum taken from healthy controls and individuals suffering from three CA subtypes: spinocerebellar ataxia type 2 (SCA2), spinocerebellar ataxia type 6 (SCA6) or ataxia-telangiectasia (AT). The proposed method classified the four groups with an accuracy of 86.3%, an impressive result for a 4-way classification. However, the confusion matrix reported by the authors indicates that no case with the SCA2 subtype was correctly classified. Because the sample size of this group (only four participants) contributed very little for the total sample size (80), it is still possible to misclassify all its cases and achieve a low error rate. In such cases, a high accuracy can be misleading, as it may reflect an overestimation of the algorithm's performance (Arbabshirani et al., 2016). Balanced accuracy, for example, is a potentially useful alternative as it calculates the average of correct predictions of each class individually (Alberg et al., 2004).

In short, since the first study published in 2013, there is already preliminary evidence that DL allows the accurate classification of a range of neurologic and psychiatric disorders, by extracting discriminating features from either single or multimodal imaging as well as other types of data such as clinical and cognitive information.

3.2. Conversion to illness

3.2.1. From Mild Cognitive Impairment to Alzheimer Dementia

A total of 8 studies have attempted to predict transition to illness using neuroimaging data, and all of them have focussed on the transition from MCI to AD (Table 2). With one exception (Liu et al., 2015a), all studies used a multimodality approach, with three of them also including clinical measures in the prognostic model. The highest accuracy (83.3%), was achieved by a model which included sMRI, PET, CSF and two clinical measures: the MMSE and the ADAS-cog (Suk et al., 2015a). Interestingly, the lowest performance (57.4%) resulted from a model which used the same input data (sMRI, PET, CSF, MMSE and ADASCog) and a similar sample size (Li et al., 2014). However, the two studies differed on the DL approach, with the former employing a semi-supervised approach with a multilayer perceptron pretrained using a stacked sparse autoencoder, and the latter using a pure supervised approach.

These findings highlight the potential impact of the DL architecture on performance, although we cannot exclude the contribution of other sample-specific factors to the results (e.g. recruitment criteria). Overall, this initial sample of studies suggests that individuals diagnosed with MCI who later convert to dementia can be identified using cutting-edge DL methods. Although, in general, accuracies are not as high as when classifying AD or MCI from healthy controls, this is not surprising since brain differences as well as clinical and cognitive symptoms between those identified as being at risk who do and do not develop a disorder are likely to be subtle. In addition to these encouraging results, the suitability of DL to multiclass classification means this analytical approach can easily be employed to examine the biomarkers of different stages of the illness. Four studies have taken advantage of this by conducting 4-way classifications to discriminate between no eminent risk of AD (healthy controls), individuals in the prodromal stage who did not (MCI-C) and did develop dementia (MCI-C) and established Alzheimer's (AD). Accuracies ranged from 46.3% to 53.8%. By using a deep Boltzmann machine to extract features from structural MRI and PET images, Liu et al. (2015a) classified the four groups with an overall accuracy of 53.8%. Suk et al. (2015b) examined the replicability of a DL approach known as deep weighted subclass-based sparse multi-task learning (DW-S2 MTL) in two different datasets, considering both binary and multi-way comparisons. The proposed model, specifically designed to mitigate the effect of less useful features for classification, showed a comparable performance for both binary (74.2% vs. 73.9%) and 4-way (53.7% vs. 47.8%) classifications, thus suggesting good replicability. Taken collectively, these studies provide initial evidence that DL methods could be used to discriminate amongst different stages of illness – a common challenge in standard clinical settings.

3.3. Treatment outcome

Prediction of response to treatment is a research area of high clinical interest. In several psychiatric and neurological disorders, a better understanding of why some patients benefit from a certain treatment whereas others do not, could help clinicians make more-effective treatment decisions and improve long-term clinical outcomes (Mechelli et al., 2015). However, so far, only one study has used DL to predict clinical response to treatment (Table 3). Munsell et al. (2015) attempted to develop an algorithm that distinguished between patients with TLE who did and did not benefit from surgical treatment. This was implemented using a stacked autoencoder to extract meaningful features from the connectome of patients who were then classified using SVM. This model, however, yielded a low accuracy of 57%. For comparison, the author investigated another option where features were extracted with an alternative linear approach instead of an autoencoder. This second model resulted in a higher accuracy of 70%. Again, this discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. This model comprised 4 layers, resulting in a high number of weights to be estimated which, together with a modest sample size (41 patients without seizures and 29 with seizures after treatment), might have resulted in overfiting.

3.4. How does DL compare to a traditional machine learning approach?

A total of twenty-five studies included in this review compared a DL model against a kernel-based model (SVM or MKL) in order to elucidate how DL compares to a more conventional ML approach. The results of these comparisons are shown in Fig. 5. It can be seen that, for the majority of studies, DL showed improved performance compared to SVM. Given the small sample of studies, it is difficult to identify specific characteristics of the studies associated with greater or smaller improvement in performance following the implementation of DL. However, a margin favouring DL studies appears to be more evident in studies that have integrated different modalities with cognitive and/or clinical data (Fig. 6). This anecdotal observation is consistent with the notion that DL is a powerful tool for detecting abstract relations within the data, especially between different types of data that are likely to be associated in complex ways, such as neuroimaging and clinical/cognitive information (Plis et al., 2014).

Since DL requires a large number of observations to learn increasingly complex patterns compared to conventional ML methods, one would expect to find a greater difference between the two methods as sample size increases. However, the effect of sample size on the difference in performance is unclear, possibly due to the small number of studies currently available. There is a minority of studies where SVM/MKL matched or even outperformed the proposed DL model. Amongst these, Munsell et al. (2015) reported the largest margin favouring SVM. However, this article had one of the smallest sample sizes (118 for the diagnostic comparison and 70 for the treatment outcome comparison) while employing one of the deepest networks with 5 layers. Notably, out of all the studies comparing the two approaches, Munsell et al. (2015) was the only one that did not make any formal attempt to prevent overfitting of the DL model, for example through the use of regularization. We note that susceptibility to overfitting becomes more pronounced when deeper and thus more complex networks are used, as in the study by Munsell et al. (2015), due to the higher number of weights to be estimated (Srivastava et al., 2014). Therefore, we speculate that the use of small sample sizes, coupled with the high-dimensionality of the data (i.e. when the number of variables highly exceeds the number of participants), may have increased the risk of overfitting in this study.

4. Discussion

ML has been gaining considerable attention in the neuroimaging community due to its advantages over traditional analytical methods based on mass-univariate statistics. In particular, ML methods take the inter-correlation between regions into account, while mass-univariate methods operate under the assumption that different regions act independently. In addition, ML methods can be used to make inferences at the single-subject level – a critical difference with mass-univariate analytical methods that are only sensitive to differences at group-level. DL is a type of ML which is increasingly used in neuroimaging after leading to major scientific advances in the areas of speech recognition, computer vision and natural language processing by significantly outperforming other state-of-the-art classification methods (Krizhevsky et al., 2012; Le et al., 2012). There are two main characteristics that distinguish DL from conventional ML methods: first, DL is capable of learning features from the raw data without the requirement for *a priori* feature selection, resulting in a more objective or less bias-prone process; second, DL uses a hierarchy of nonlinear transformations, which make this approach ideally suited for detecting complex, scattered and subtle patterns in the data. Given its ability to detect abstract patterns from the data, DL can be considered a promising tool in neuroimaging, as most brain-based disorders are characterised by a scattered and diffused pattern of neuroanatomical and neurofunctional alterations (Plis et al., 2014). In previous sections of this review, we have described the most common DL architectures and have provided an overview of the studies that have applied DL to neuroimaging data to investigate psychiatric and neurological disorders. In this final section, we discuss the main themes that have emerged from the review of these studies. These will include (i)

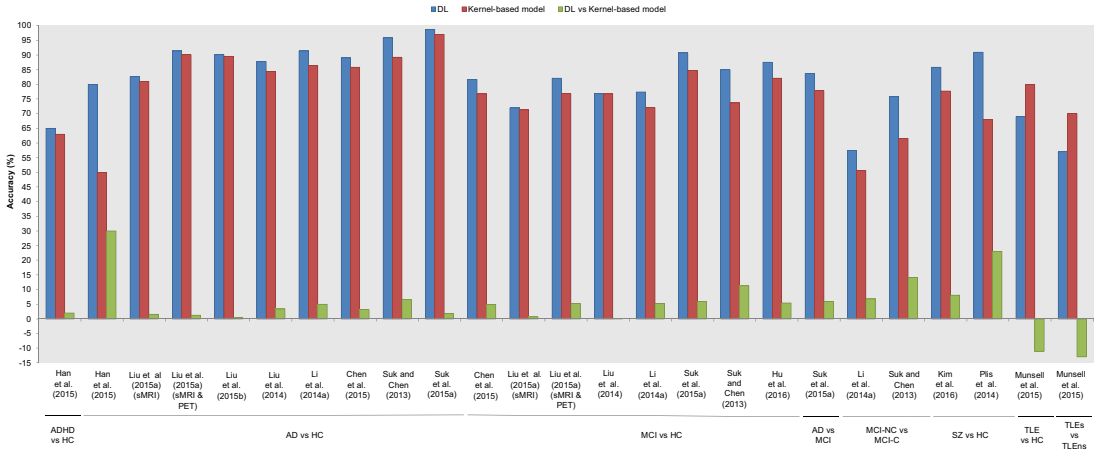


Fig. 5. Results of studies comparing DL and kernel-based models. The graph shows the accuracies (F-score for Plis et al., 2014) for DL models (blue), kernel-based models (red) and the difference between the two (green). HC, healthy controls; ADHD, attention deficit and hyperactive disorder; AD, Alzheimer's disease; MCI, mild cognitive impairment; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; SZ, schizophrenia; TLE, temporal lobe epilepsy; TLEs, temporal lobe epilepsy with seizures after treatment; TLEns, temporal lobe epilepsy without seizures after treatment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

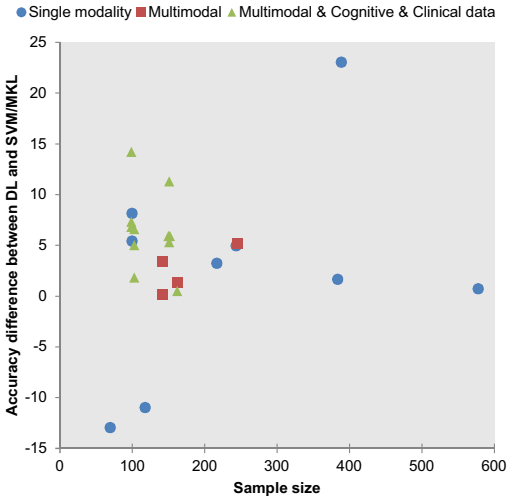


Fig. 6. Difference in performance of DL against kernel-based methods for single modality, multimodal as well as for multimodal with cognitive/clinical data studies, according to sample size.

consistencies and inconsistencies in the existing literature (ii) the promise of CNNs, (iii) the issue of multiclass classification, (iv) how DL performs compared with conventional ML methods, (v) interpretability of DL in neuroimaging, (vi) the challenge of overfitting and (vii) technical expertise and computational requirements. We conclude by discussing possible directions for future research.

4.1. Main conclusions from the existing literature

The majority of published studies have been conducted in patients with MCI and/or AD; this may be explained by the availability of ADNI, a very large open-source dataset including thousands of patients, to the neuroimaging community (Mueller et al., 2005a, 2005b). However, studies have also been conducted in other disorders including ADHD, psychosis, TLE and cerebellar ataxia. Taken collectively, the findings published so far suggest that DL can be applied to neuroimaging data, including both structural and functional modalities, to classify diagnostic groups from healthy individuals. Indeed, the performance of the classifiers has been consistently high, with several studies reporting accuracies above 95% for binary classifications between patients and controls (Deshpande et al., 2015; Hosseini-Asl et al., 2016; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Suk and Shen, 2013; Suk et al., 2015a; Suk et al., 2015b). Nevertheless, the application of a supervised model for diagnostic classification is arguably circular: since diagnostic labels in the training and testing datasets are predetermined through clinical examination, logic dictates that a perfect performance from an ML algorithm will simply mimic clinical assessment. Being able to predict a future diagnosis, or anticipate who will and will not benefit from a certain treatment, are questions of greater translational value in clinical practice. A total of 8 studies have applied DL to neuroimaging data acquired from individuals with MCI to predict subsequent transition to AD with promising results. For example, Suk et al. (2015a) successfully predicted conversion from MCI to AD with 83.3% accuracy, after combining structural MRI and PET data. However, no studies have yet examined transition to illness in other psychiatric disorders with a prodromal phase, such as psychosis, even though we know that it is possible to distinguish between converters and non-converters using conventional ML (Zarogianni et al., 2013; Pettersson-Yeo et al., 2013; Valli et al., 2016). To our knowledge only one study has used DL to predict treatment outcome. Munsell et al. (2015) achieved an accuracy of 57% when classifying TLE patients who did and did not suffer from seizures after surgical intervention. As discussed earlier, however, this modest result could potentially be explained by the absence of formal strategies to avoid overfitting of the DL model.

DL is a very flexible approach, meaning that is it possible to combine different architectures and manipulate a range of hyperparameters within the same model. In addition, the vast majority

of existing studies have been published in the last 2 years, and therefore the field of DL applied to neuroimaging of brain-disorders should be considered still at a very early stage. Possibly as a result of this combination of flexibility and novelty, the methodology of the studies reviewed in this article varied considerably. For example, some studies employed a whole-brain approach whereas others focussed on a subset of regions of interest; some studies used the raw data without any form of feature selection whereas others performed a number of transformations on the data to select relevant features; and different studies used different DL architectures. Such methodological variability means that, at present, the reliability and replicability of the existing results remain unclear.

4.2. The promise of convolutional neural networks

CNNs are a particular type of feedforward neural network inspired by how the human visual cortex process information. Over the past decade, CNNs have been breaking records in computer vision across several competitions, making this approach a very promising one (Krizhevsky et al., 2012). Consistent with this, our review has shown that CNNs have generated the most encouraging results in the context of neuroimaging. In its raw form, neuroimaging data comprises millions of voxels. Considering the current computational resources available, putting all voxel intensities through a fully connected network would lead to an unfeasible number of weights to be estimated. Two intrinsic properties of CNNs – weight sharing and local connectivity – result in a significantly reduced number of weights, making it computationally possible to run the network at the voxel-level. Although in neuroimaging CNNs have only been used to examine MCI and AD patients, the accuracies of the studies published so far have been consistently high (i.e. ≥95% for AD and ≥86% for MCI versus controls). High accuracies have been observed with different modalities including structural MRI (Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan and Montana, 2015), resting-state fMRI (Sarraf and Tofighi, 2016) and CT imaging (Gao and Hui, 2016), as well as with small (Gao and Hui, 2016; Sarraf and Tofighi, 2016) and large (Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan and Montana, 2015) sample sizes. Hosseini-Asl et al. (2016) used an alternative and interesting approach which involved pre-training a CNN in one Alzheimer's dataset (CADDementia) and then fine-tuning and testing it in another dataset from the same diagnostic group (ADNI). The results were very promising for both 2-way and 3-way classifications (HC vs. AD; HC vs. MCI; AD vs. MCI; and HC vs. AD vs. MCI), although it should be noted that the ADNI sample was of modest size. Taken together, these results are in line with the successful performances of CNN-based models reported in other scientific areas, and highlight CNNs as a promising tool in neuroimaging.

4.3. From binary to multiclass classifications

In the context of neuroimaging, the vast majority of conventional ML studies have relied on binary classifications involving the comparison between a group of patients and a group of healthy controls (Orrù et al., 2012; Wolfers et al., 2015). This can be explained by the fact that these studies have typically employed SVM, which was originally designed for binary classification problems (Hsu and Lin, 2002). However, the real challenge for clinicians is not to differentiate between patients and controls but to develop biomarkers which could be used to choose amongst alternative diagnoses or different stages of illness progression. Looking forward, therefore, ML models will need to be able to discriminate amongst several possible alternatives in order to inform real-world clinical decision making. Many approaches have been proposed to enable SVM to handle multiclass classification problems (Fei and Liu, 2006; Hsu and Lin, 2002). However, this is still an active research area (Kumar and Gopal, 2011) and none of the proposed approaches have been tested in the context of neuroimaging. Most neuroimaging studies using SVM addressed the multiclass problem by performing several binary classifications (for example, AD vs. HC, MCI vs. HC and AD vs. MCI) or one-against-all classifications (for example, AD vs. MCI & HC and MCI vs. AD & HC). DL however, requires less technical effort to perform multiclass comparisons, and therefore could provide a solution to this issue. This is mainly due to the use of the so-called softmax function in the output layer, which can be considered an extension of the binary logistic regression to several classes. Here the output reflects the probability of belonging to each class, which is a more intuitive index of class membership than some of the most sophisticated indices being developed for SVM multiclass solutions (Fei and Liu, 2006). In light of its suitability for multiclass classification, a number of studies have used DL to carry out 3 or 4-way classifications between different disorder subtypes or different stages of illness. For example, three of these studies were able to classify children into healthy controls and three ADHD subtypes (inattentive, hyperactive and combined) (Hao et al., 2015; Kuang and He, 2014; Kuang et al., 2014). Notably, there is also preliminary evidence for the use of DL to distinguish between individuals at no imminent risk of dementia, those identified at risk who will and will not develop dementia, and those with established Alzheimer's disease (Liu et al., 2015a; Liu et al., 2014; Suk et al., 2015b). These are encouraging findings, as they highlight how DL could help bridge the existing gap between neuroimaging findings and real-world clinical practice.

4.4. Is deep learning superior to conventional machine learning?

Despite the success of DL in several scientific areas, the superiority of this analytical approach in neuroimaging is yet to be demonstrated. On the one hand, DL has been described as a potentially more powerful approach than conventional shallow ML, as it is capable of learning highly intricate and abstract patterns from the data, which can particularly useful in the case of brain-based disorders (Plis et al., 2014). On the other hand, given that neuroimaging data is very high-dimensional, the nonlinear approach of DL might not be advantageous as there are not enough data points to extract meaningful nonlinear patterns from the data, whereas the linear approach employed in conventional shallow ML might be more appropriate. Here we tried to clarify this issue by systematically examining the difference in performance between DL and conventional shallow ML in studies which used both approaches. A total of twenty-five studies reported classification accuracy for both DL and conventional shallow ML, with the latter being a kernel-based method, either SVM or MKL. For the majority of these studies DL performed better than conventional shallow ML as shown in Fig. 5, and in some cases the difference was by a reasonable margin (e.g. Han et al., 2015; Plis et al., 2014; Suk and Chen, 2013).

From the available evidence, it is not clear whether DL tends to perform better under specific circumstances, for example depending on the modality type or the sample size. However, our systematic review provides anecdotal evidence that studies combining imaging and non-imaging data tend to have a larger margin in favour of DL (see Fig. 6). This is consistent with the notion that the association between brain abnormalities and cognitive symptoms, for example, is likely to exist at a deep and abstract level, and as such can be captured more effectively by DL methods than traditional shallow ML methods (Plis et al., 2014).

We know that the application of traditional shallow ML methods to neuroimaging data leads to higher and more stable accuracies as the sample size increases (Nieuwenhuis et al., 2012). One would expect this to be especially true for DL: since a deep model is inherently more complex than conventional shallow ML models, larger

sample sizes should be needed to compensate for the greater number of parameters to be estimated and to take full advantage of DL's ability to detect highly intricate and abstract patterns in the data. We were therefore expecting to see an increase in the margin by which DL outperforms kernel-based methods as sample sizes increase. Such increase however was not observed, as the pattern of difference in performance did not seem to vary systematically with sample size; one possibility is that larger sample sizes than those used in the existing literature would be required to detect increases in the margin by which DL outperforms kernel-based methods.

In conclusion, our review suggests that, overall, DL performs better than conventional shallow ML. In light of the increasing interest in DL, however, we cannot exclude a publication bias which favoured studies showing the superiority of this new analytical approach relative to conventional shallow ML methods (Boulesteix et al., 2013). As the number of studies applying DL to neuroimaging data increases, a thorough assessment of publication bias would be useful to establish the reliability of this initial trend in favour of DL.

4.5. Interpretability of DL in neuroimaging

Despite having demonstrated state-of-the-art performances across several fields, DL has been under scrutiny for its lack of transparency during the learning and testing processes (Alain and Bengio, 2016; Lou et al., 2012; Yosinski et al., 2015). For example, deep neural networks have been referred to as a "black box" in contrast with other techniques, such as logistic regression, which are less complex and more intuitive. Such lack of transparency has important implications for the interpretability of the results when DL is applied to neuroimaging data. Due to the multiple nonlinearities, it can be challenging to trace the consecutive layers of weights back to the original brain image in order to identify which features (e.g. regions) are providing the greatest contribution to classification (Suk et al., 2015a). This information however would be useful in the context of clinical neuroimaging where the aim is not only to detect but also localise abnormalities. A first potential issue is that a model with an excellent performance may be using irrelevant features (e.g. orientation of the images, imaging artefacts), as oppose to clinically meaningful information (e.g. regional grey matter, connectivity between different brain regions), to classify participants. A second potential issue is that an accurate model which provides no information about the underlying neuroanatomical or neurofunctional alterations would be of limited clinical utility, for example with respect to treatment development and optimization.

Despite its complex inner workings which make the visualization and interpretation of the weights challenging, DL *can* be used in a way which enables transparency. This is illustrated by several neuroimaging studies included in this review that did report the most important features (e.g., Deshpande et al., 2015; Kim et al., 2016; Liu et al., 2014; Suk et al., 2016). However, these studies used a variety of approaches to isolate the most informative features, and at present there is no standard and intuitive method for visualizing weights or interpreting latent feature representations (Suk et al., 2015a). This has motivated several attempts to develop new and intuitive ways of enhancing the interpretability of DL within the recent literature (e.g., Grün et al., 2016; Samek et al., 2015; Simonyan et al., 2013; Yosinski et al., 2015; Zeiler and Fergus, 2014). There are two main methodological approaches to address this issue, including input modification methods and deconvolution methods. Input modification methods are visualization techniques that involve the systematic modification of the input and the measurement of any resulting changes in the output as well as in the activation of the artificial neurons in the intermediate layers of the network. An example of these methods is the so-called occlusion method (Zeiler and Fergus, 2013) which involves covering portions of the input image up to find the areas of the input data that influence the probability of the output classes. In contrast, deconvolution methods aim to determine the contribution of one or more features of the input data to the output. This involves selecting an activation of interest in an output neuron and then computing the contribution of each neuron in the next lower layers to this activation. Here a number of strategies are available to model the nonlinearities present across the layers, for example, deconvnet (Zeiler and Fergus, 2013) and guided backpropagation (Springenberg et al., 2014).

4.6. The challenge of overfitting

Overfitting is arguably one of the main challenges in ML. Given their inherent complexity, DL networks are particularly prone to overfitting, i.e., learning irrelevant fluctuations in the data that limit generalizability. Not surprisingly, different approaches to address this issue, known as regularization strategies, have been developed and are now present in most DL algorithms. In section 2.1.4 we described some of the most commonly used regularization strategies applied to modern DL, namely weight decays and dropout. As expected, several studies reviewed here have used some form of regularization. The majority (e.g., Hosseini-Asl et al., 2016; Kim et al., 2016; Liu et al., 2015a) have employed the L1 or L2 norms, which prevent overfitting by penalizing very low or very high weight values. At least one study (Li et al., 2014) employed dropout, where a random number of nodes and respective connections are temporarily removed to extract different sets of features that can independently produce a useful output. The importance of regularization strategies in DL could potentially account for the fact that Munsell and colleagues, who trained 4- and 5-hidden layer models (for inferring diagnostic and treatment outcome, respectively) without using any form of regularization, reported such low performance for DL (Munsell et al., 2015).

An additional approach for minimising the risk of overfitting involves reducing the dimensionality of the data before inputting them into the model. A possible way of achieving this is by extracting region- or patch-level features (as opposed to using voxel-level data). Using different types of features (whether voxel, patch or region) can have implications for how detailed the information inputted into the model is (for example, voxel-level features are very detailed, and also very noisy; region-level features on the other hand, ignore more localized patterns and are less sensitivity to noise). Another option to reduce dimensionality is feature selection. Feature selection is common in conventional ML, where linear methods such as principal component analysis, independent component analysis or elastic net, are used to select the most discriminating features that are then fed to a classifier. However, the use of conventional feature selection methods prior to a DL model seems counterintuitive, since one of the main advantages of DL is the ability to learn, through a purely data-driven method, the most useful features for classification. Several studies reported in this review have attempted to reduce the dimensionality of the data by extracting region- or patch-level features, using feature selection, or combining the two approaches. We note, however, that all CNN-based models were applied to voxel-level data without being preceded by any form of feature selection and yet reported consistently high performances on unseen data. This suggests that DL, and CNN-models and particular, can perform well with neuroimaging data without the requirement to downsize or even preprocess the data. For example, Hosseini-Asl et al. (2016) achieved high levels of accuracy after applying a CNN to voxel-level data without any preprocessing or even skull stripping of the images. This finding has potential implications for the development of clinical tools, as it suggests that it might be possible to apply DL to raw neuroimaging data, thereby saving time as well as technical resources.

4.7. Technical expertise and computational requirements

The studies reviewed in this article employed a wide range of DL architectures and hyperparameters. Such flexibility is what makes DL a very powerful tool but comes at a potentially high cost. The number of layers, the number of nodes within each layer and the activation function of each node are only a few examples of a long list of variables one has to consider when designing and optimizing a DL model. Automated optimization strategies are not yet widely available, making optimisation a manual process that requires a great deal of technical expertise and is potentially prone to subjective bias. Since the number of parameters to be estimated is very large, the computational requirements of DL are also more demanding than those of conventional ML methods. For example, Kim et al. (2016) reported that the estimation of a DL model with three hidden layers took 100 times longer than the estimation of a standard SVM model (~3.3 days vs. 0.8 h). However, with the fast-growing availability of graphical processing units (GPUs), the application of DL to neuroimaging data is likely to become less and less time-consuming in the future.

5. Conclusions and future directions

While still in its initial stages, the application of DL in neuroimaging has shown promising results and has the potential of leading to fundamental advances in the search for imaging-based biomarkers of psychiatric and neurologic disorders. Nevertheless, several improvements will be required before the full potential of DL in neuroimaging can be achieved. Firstly, given the complexity of DL models, we need to move away from studies with small to modest sample sizes in favour of much larger cohorts. A possible way of achieving this is through multi-centre collaborations, in which data is collected using the same recruitment criteria and scanning protocols across sites. A further way of increasing the sample size is through multi-site data sharing initiatives, such as ADNI for Alzheimer's disease and ADHD-200 for ADHD. Secondly, the integration of CNN and recurrent neural networks (i.e. networks that allow the processing of data with sequential inputs such as videos or speech) is likely to lead to significant advances in DL in the next few years (Donahue et al., 2015). In neuroimaging, this integration could be particularly useful for analysing fMRI data, as it would allow the detection of intricate spatial patterns while simultaneously modelling the temporal component of the BOLD signal. Thirdly, we anticipate that an increasing number of neuroimaging studies will make use of transfer learning, which involves using previously learned features from a large sample of similar enough images. This could help tackle the curse of dimensionality – a common problem in neuroimaging studies of brain disorders (Gupta et al., 2013; Hosseini-Asl et al., 2016). Evidence from vision science, where deeper models such as VGG net (Simonyan and Zisserman, 2014), residuals networks (He et al., 2015) and Inception-v4 (Szegedy et al., 2016) are achieving the highest performances, suggests that transfer learning could be particularly useful when deeper models are employed. Fourthly, we suggest that the so-called augmentation technique – which it is commonly used in computer vision – could be useful in the context of neuroimaging. This technique involves increasing the sample size by applying transformations to the data (e.g., rotation, shear, scaling), and then train a model that is invariant to such transformations. The use of augmentation could also address the issue of modest sample sizes and lead to a decrease in prepossessing time (because steps such as rotation may become redundant). Finally, the use of DL to predict continuous scores is another interesting area for further research with potential clinical applicability, following the encouraging results obtained using conventional ML methods (e.g. Gong et al., 2014; Stonnington et al., 2010; Tognin et al., 2014). So far, only one study has used DL to predict clinical scores from structural MRI scans in patients with Alzheimer's disease (Brosch and Tam, 2013).

In conclusion, the capacity of DL models to learn complex and abstract representations through nonlinear transformations, makes this a promising approach to single subject prediction in neuroimaging. While there are still important challenges to overcome, the findings reviewed here provide preliminary evidence supporting the potential role of DL in the future development of diagnostic and prognostic biomarkers of psychiatric and neurologic disorders.

Acknowledgements

Sandra Vieira is supported by a PhD studentship from the Fundação para a Ciência e a Tecnologia (FCT), research grant SFRH/BD/103907/2014. Walter H.L. Pinaya gratefully acknowledges support from FAPESP (Brazil), grant #2013/05168-7, São Paulo Research Foundation. Andrea Mechelli is supported by the Medical Research Council (ID99859).

References

Alain, G., Bengio, Y., 2016. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.
Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of overall accuracy to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460–465.
Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2016. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage, 137–165.
Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends® Mach. Learn. 2, 1–127.
Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. Adv. Neural Inf. Process. Syst., 2546–2554.
Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kotter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. 107, 4734–4739.
Boulesteix, A.L., Lauer, S., Eugster, M.J., 2013. A plea for neutral comparison studies in computational sciences. PLoS One 8, e61562.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Proceedings of the IEEE 20th International Conference on Pattern Recognition, 3121–3124.
Brosch T., Tam R., Alzheimer's Disease Neuroimaging Initiative, 2013. Manifold learning of brain MRIs by deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 633–640. Springer Berlin Heidelberg.
Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V.D., Dwyer, D.B., von Saldern, S., Urquijo, M.F., Falkai, P., Koutsouleris, N., 2016. Classifying schizophrenia using multimodal multivariate pattern recognition analysis: evaluating the impact of individual clinical profiles on the neurodiagnostic performance. Schizophr. Bull. 42, S110–S117.
Calhoun, V.D., Sui, J., 2016. Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. Biol. Psychiatry: Cogn. Neurosci.Neuroimag. 1, 230–244.
Chen, Y., Shi, B., Smith, C.D., Liu, J., 2015. Nonlinear Feature Transformation and Deep Fusion for Alzheimer's Disease Staging Analysis. In: International Workshop on Machine Learning in Medical Imaging, 304–312. Springer International Publishing.
Deshpande, G., Wang, P., Rangaprakash, D., Wilamowski, B., 2015. Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. IEEE Trans. Cybernet. 45, 2668–2679.
Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2625–2634.
Fei, B., Liu, J., 2006. Binary tree of SVM: a new fast multiclass training and classification algorithm. IEEE Trans. Neural Netw. 17, 696–704.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U. S. A. 102, 9673–9678.
Gao, X.W., Hui, R., 2016. A deep learning based approach to classification of CT brain images. In: Science and Information Conference, London, UK.
Gelbart, M.A., Snoek, J., Adams, R.P., 2014. Bayesian optimization with unknown constraints. arXiv preprint arXiv:1403.5607.
Gong, Q., Li, L., Du, M., Pettersson-Yeo, W., Crossley, N., Yang, X., Li, J., Huang, X., Mechelli, A., 2014. Quantitative prediction of individual psychopathology in trauma survivors using resting-state FMRI. Neuropsychopharmacology 39, 681–687.
Grün, F., Rupprecht, C., Navab, N., Tombari, F., 2016. A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks. arXiv preprint arXiv:1606.07757.
Gupta, A., Ayhan, M., Maida, A., 2013. Natural image bases to represent neuroimaging data. International Conference on Machine Learning, 987–994.
Han X., Zhong Y., He L., Philip S.Y., Zhang L., 2015. The unsupervised hierarchical convolutional sparse auto-encoder for neuroimaging data classification. In: International Conference on Brain Informatics and Health, 156–166. Springer International Publishing.
Hao, A.J., He, B.L., Yin, C.H., 2015. Discrimination of ADHD children based on deep bayesian network. 2015 International Conference on Biomedical Image and Signal Processing, 1–6.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York, NY.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.
Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Hosseini-Asl, E., Gimel'farb, G., El-Baz, A., 2016. Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network. arXiv preprint arXiv:1607.00556.
Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.
Hu, C., Ju, R., Shen, Y., Zhou, P., Li, Q., 2016. Clinical decision support for Alzheimer's disease based on deep learning and brain network. Proceedings of the IEEE International Conference on Communications, 1–6.
Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., Handwerker, D.A., Keilholz, S., Kiviniemi, V., Leopold, D.A., de Pasquale, F., Sporns, O., Walter, M., Chang, C., 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360–378.
Kennedy, D.P., Courchesne, E., 2008. The intrinsic functional organization of the brain is altered in autism. Neuroimage 39, 1877–1885.
Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage 124, 127–146.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems., pp. 1097–1105.
Kuang, D., He, L., 2014. Classification on ADHD with deep learning. Proceedings of the International Conference on Cloud Computing and Big Data, 27–32.
Kuang, D., Guo, X., An, X., Zhao, Y., He, L., 2014. Discrimination of ADHD based on fMRI data with deep belief network. International Conference on Intelligent Computing, 225–232.
Kumar, M.A., Gopal, M., 2011. Reduced one-against-all method for multiclass SVM classification. Expert Syst. Appl. 38, 14238–14248.
Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation. Proceedings of the 24th International Conference on Machine Learning, 473–480.
Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Ng, A., 2012. Building high-level features using large scale unsupervised learning. International Conference on Machine Learning 103.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Li, F., Tran, L., Thung, K.H., Ji, S., Shen, D., Li, J., 2014. Robust deep learning for improved classification of AD/MCI patients. International Workshop on Machine Learning in Medical Imaging, 240–247.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's Disease with deep learning. IEEE 11th International Symposium on Biomedical Imaging, 1015–1018.
Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., 2015a. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. 62, 1132–1140.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D.D., 2015b. Multi-phase feature representation learning for neurodegenerative disease diagnosis. Australasian Conference on Artificial Life and Computational Intelligence, 350–359.
McCulloch, W., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 7, 115–133.
Mechelli, A., Prata, D., Kefford, C., Kapur, S., 2015. Predicting clinical response in people at ultra-high risk of psychosis: a systematic and quantitative review. Drug Discovery Today 20, 924–927.
Milham, M.P., Fair, D., Mennes, M., Mostofsky, S.H., 2012. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Front. Syst. Neurosci. 6, 62.
Moody, J., Hanson, S., Krogh, A., Hertz, J.A., 1995. A simple weight decay can improve generalization. Adv. Neural Inf. Process. Syst. 4, 950–957.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Alzheimer's disease neuroimaging initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005a. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dementia 1, 55–66.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005b. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877.
Mulders, P.C., van Eijndhoven, P.F., Schene, A.H., Beckmann, C.F., Tendolkar, I., 2015. Resting-state functional connectivity in major depressive disorder: a review. Neurosci. Biobehav. Rev. 56, 330–344.
Munsell, B.C., Wee, C.Y., Keller, S.S., Weber, B., Elger, C., da Silva, L.A.T., Nesland, T., Styner, M., Shen, D., Bonilha, L., 2015. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. Neuroimage 118, 219–230.
Nieuwenhuis, M., van Haren, N.E., Pol, H.E.H., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage 61, 606–612.
Nowlan, S.J., Hinton, G.E., 1992. Simplifying neural networks by soft weight-sharing. Neural Comput. 4, 473–493.
Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36, 1140–1152.
Page, A., Turner, J.T., Mohsenin, T., Oates, T., 2014. Comparing raw data and feature extraction for seizure detection with deep learning methods. International Florida Artificial Intelligence Research Society Conference.
Payan, A., Montana, G., 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint arXiv: 1502.02506.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Pettersson-Yeo, W., Benetti, S., Marquand, A.F., Dell'Acqua, F., Williams, S.C.R., Allen, P., Prata, D., McGuire, P., Mechelli, A., 2013. Using genetic: cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. Psychol. Med. 43, 2547–2562.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 1–11.
Radua, J., Borgwardt, S., Crescini, A., Mataix-Cols, D., Meyer-Lindenberg, A., McGuire, P.K., Fusar-Poli, P., 2012. Multimodal meta-analysis of structural and functional brain changes in first episode psychosis and the effects of antipsychotic medication. Neurosci. Biobehav. Rev. 36, 2325–2333.
Samek, W., Binder, A., Montavon, G., Bach, S., Müller, K.R., 2015. Evaluating the visualization of what a deep neural network has learned. arXiv preprint arXiv:1509.06321.
Sarraf, S., Tofighi, G., 2016. Classification of Alzheimer's Disease using fMRI Data and Deep Learning Convolutional Neural Networks. arXiv preprint arXiv:1603.08631.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.
Schultz, C.C., Fusar-Poli, P., Wagner, G., Koch, K., Schachtzabel, C., Gruber, O., Sauer, H., Schlösser, R.G., 2012. Multimodal functional and structural imaging investigations in psychosis research. Eur. Arch. Psychiatry Clin. Neurosci. 262, 97–106.
Sheffield, J.M., Barch, D.M., 2016. Cognition and resting-state functional connectivity in schizophrenia. Neurosci. Biobehav. Rev. 61, 108–120.
Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
Springenberg, J.T., Dosovitskiy, A., Brox, T., and Riedmiller, M., 2014. Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806.
Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S., 2010. Alzheimer Disease Neuroimaging Initiative. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51, 1405–1413.
Suk, H.I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. International Conference on Medical Image Computing and Computer-Assisted Intervention, 583–590.
Suk, H.I., Lee, S.W., Shen, D., 2014. Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101, 569–582.

Suk, H.I., Lee, S.W., Shen, D., 2015a. Alzheimer's disease neuroimaging initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859.
Suk, H.I., Lee, S.W., Shen, D., 2015b. Alzheimer's Disease Neuroimaging Initiative. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. Brain Struct. Funct., 1–19.
Suk, H.I., Wee, C.Y., Lee, S.W., Shen, D., 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. Neuroimage 129, 292–307.
Szegedy, C., Ioffe, S., Vanhoucke, V., 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261.
Tognin, S., Pettersson-Yeo, W., Valli, I., Hutton, C., Woolley, J., Allen, P., McGuire, P., Mechelli, A., 2014. Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis. Front. Psychiatry 4, 187.
Valli, I., Marquand, A.F., Mechelli, A., Raffin, M., Allen, P., Seal, M.L., McGuire, P., 2016. Identifying individuals at high risk of psychosis: predictive utility of Support Vector Machine using structural and functional MRI data. Front. Psychiatry 7.
van der Meer, L., Costafreda, S., Aleman, A., David, A.S., 2010. Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. Neurosci. Biobehav. Rev. 34 (6), 935–946.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.
Willette, A.A., Calhoun, V.D., Egan, J.M., Kapogiannis, D., 2014. Alzheimer' s Disease Neuroimaging Initiative. Prognostic classification of mild cognitive impairment and Alzheimer's disease: MRI independent component analysis. Psychiatry Res.: Neuroimag. 224, 81–88.
Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349.
Yang, Z., Zhong, S., Carass, A., Ying, S.H., Prince, J.L., 2014. Deep learning for cerebellar ataxia classification and functional score regression. International Workshop on Machine Learning in Medical Imaging, 68–76.
Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.
Yung, A.R., Yuen, H.P., McGorry, P.D., Phillips, L.J., Kelly, D., Dell'Olio, M., Francey, S.M., Cosgrave, E.M., Killackey, E., Stanford, C., Godfrey, K., Buckby, J., 2005. Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. Aust. N. Z. J. Psychiatry 39, 964–971.
Zarogianni, E., Moorhead, T.W., Lawrie, S.M., 2013. Towards the identification of imaging biomarkers in schizophrenia: using multivariate pattern classification at a single-subject level. NeuroImage: Clin. 3, 279–289.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In European Conference on Computer Vision, 818–833. Springer International Publishing.
Zhang, D., Shen, D., 2012. Alzheimer's Disease Neuroimaging Initiative. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7, e33182.

Schizophrenia Bulletin
doi:10.1093/schbul/sby189

Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence

Sandra Vieira[1], Qi-yong Gong[*,2,3], Walter H. L. Pinaya[1,4], Cristina Scarpazza[1,5], Stefania Tognin[1], Benedicto Crespo-Facorro[6,7], Diana Tordesillas-Gutierrez[6,8], Victor Ortiz-García[6,7], Esther Setien-Suero[6,7], Floortje E. Scheepers[9], Neeltje E. M. van Haren[10], Tiago R. Marques[1], Robin M. Murray[1], Anthony David[1], Paola Dazzan[1], Philip McGuire[1], and Andrea Mechelli[1,]

[1]Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom; [2]Huaxi MR Research Center (HMRRC), Department of Radiology, West China Hospital of Sichuan University, Chengdu, China; [3]Department of Psychoradiology, Chengdu Mental Health Center, Chengdu, China; [4]Centre of Mathematics, Computation, and Cognition, Universidade Federal do ABC, São Paulo, Brazil; [5]Department of General Psychology, University of Padova, Padova, Italy; [6]Centro Investigación Biomédica en Red de Salud Mental (CIBERSAM), Spain; [7]Department of Psychiatry, University Hospital Marqués de Valdecilla, School of Medicine, University of Cantabria-IDIVAL, Santander, Spain; [8]Neuroimaging Unit, Technological Facilities, Valdecilla Biomedical Research Institute IDIVAL, Santander, Cantabria, Spain; [9]Department of Psychiatry, University Medical Centre Utrecht, Utrecht, The Netherlands; [10]Brain Centre Rudolf Magnus, University Medical Centre Utrecht, Utrecht, The Netherlands

*To whom correspondence should be addressed; Huaxi MR Research Center (HMRRC), Department of Radiology, West China Hospital of Sichuan University, Chengdu 610041, China; tel: +86(0) 28 8542 3503, e-mail: qiyonggong@hmrrc.org.cn

Despite the high level of interest in the use of machine learning (ML) and neuroimaging to detect psychosis at the individual level, the reliability of the findings is unclear due to potential methodological issues that may have inflated the existing literature. This study aimed to elucidate the extent to which the application of ML to neuroanatomical data allows detection of first episode psychosis (FEP), while putting in place methodological precautions to avoid overoptimistic results. We tested both traditional ML and an emerging approach known as deep learning (DL) using 3 feature sets of interest: (1) surface-based regional volumes and cortical thickness, (2) voxel-based gray matter volume (GMV) and (3) voxel-based cortical thickness (VBCT). To assess the reliability of the findings, we repeated all analyses in 5 independent datasets, totaling 956 participants (514 FEP and 444 within-site matched controls). The performance was assessed via nested cross-validation (CV) and cross-site CV. Accuracies ranged from 50% to 70% for surfaced-based features; from 50% to 63% for GMV; and from 51% to 68% for VBCT. The best accuracies (70%) were achieved when DL was applied to surface-based features; however, these models generalized poorly to other sites. Findings from this study suggest that, when methodological precautions are adopted to avoid overoptimistic results, detection of individuals in the early stages of psychosis is more challenging than originally thought. In light of this, we argue that the current evidence for the diagnostic value of ML and structural neuroimaging should be reconsidered toward a more cautious interpretation.

Keywords: multivariate pattern recognition/classification/psychosis/neuroimaging/multi-site

Introduction

Over the last 3 decades, traditional mass-univariate neuroimaging approaches have revealed neuroanatomical abnormalities in individuals with psychosis.[1–5] Because these abnormalities were detected using group-level inferences, it has not been possible to use this information to make diagnostic and treatment decisions about individual patients. Machine learning (ML) is an area of artificial intelligence that promises to overcome this issue by learning meaningful patterns from the imaging data and using this information to make predictions about unseen individuals.[6] Several ML studies have attempted to use neuroanatomical data to distinguish patients with established schizophrenia from healthy individuals, with promising results.[7–10] At present, however, there are two important limitations in the existing literature that limit the translational applicability of the findings in real-world clinical practice. First, given the well-established effects of illness chronicity and antipsychotic medication on brain structure,[11–15] it is unclear to what extent classification was based on neuroanatomical changes associated with these factors rather than the onset of the illness per se. Consistent with this, both disease-stage and antipsychotic medication were identified as significant

moderators in a recent meta-analysis of diagnostic biomarkers in schizophrenia.[7] Also in line with this, Pinaya et al[16] reported that the same ML model that was able to distinguish between patients with established schizophrenia and healthy controls (HCs) with an accuracy of 74% showed poor generalizability (56%) when applied to a cohort of individuals with first episode psychosis (FEP). Taken collectively, these findings suggest that representations learned from patients with established schizophrenia may not be applicable to individuals with a first episode of the illness. Second, the clinical utility of any ML-based diagnostic tool for detecting patients with an established illness is likely to be very limited; in contrast, detecting the initial stages of an illness, when diagnosis may be uncertain and treatment is yet to be decided, is likely to have much greater clinical utility.

So far only a limited number of studies have applied ML to neuroanatomical data in the initial stages of the illness when the effects of illness chronicity and antipsychotic medication are minimal. These studies have produced inconsistent results, including poor (eg, 51% in Winterburn et al[17]), modest (eg, 63% in Pettersson-Yeo et al[18]), and good (eg, 86% in Borgwardt et al[19] or 85% in Xiao et al[20]) accuracies. There are a number of possible reasons for such inconsistency. First, most of the studies used small samples ($N \leq 50$) (see Kambeitz et al[7] for a meta-analysis), which have been shown to yield unstable results.[21,22] Second, the vast majority of studies used data from a single site, and as such may have generated results that were specific to the characteristic of the local sample rather than the illness per se. Third, a series of recent articles have highlighted potential methodological issues that may have caused inflated results in some of the published studies.[9,17,22–25] These issues include, eg, (1) failure to use a nested cross-validation (CV) framework to avoid *knowledge-leakage* between training and test sets; (2) failure to perform feature transformation and/or selection within a rigorous CV framework resulting in so-called "double dipping"; (3) publication bias leading to an overrepresentation of positive findings, especially in studies with small samples and (4) failure to test performance on additional independent samples. Also, we note that all studies have employed traditional "shallow" ML techniques, such as support vector machine and logistic regression. The intuitiveness of such techniques has made them very popular in neuroimaging studies of psychiatric and neurological disease. Deep learning (DL) is an alternative type of ML, which has been gaining considerable attention in clinical neuroimaging.[9,16,23,26] Contrary to traditional ML, where the immediate input data are used to extract patterns (hence the term "shallow"), DL learns complex latent features of brain structure through consecutive nonlinear transformations (hence the term "deep"), which are then used for classification. Given its ability to learn more intricate and abstract patterns, DL might be particularly suitable to detect the subtle and heterogeneous neuroanatomical abnormalities characteristic of the early stages of psychosis.[1,27,28]

This study aims to elucidate the extent to which the application of ML to neuroanatomical data allows distinction between patients with FEP and HCs at the individual level. To overcome the limitations of previous studies, we used a total of 5 datasets from different sites, each with a sample size above the recommended threshold for a stable performance,[21] and employed both shallow and deep ML techniques. In addition, following a series of recent articles highlighting potential methodological issues in the existing literature,[9,17,22–25] we put in place a series of precautions to minimize the risk of overfitting. On the basis of previous studies, we hypothesize that (1) FEP and HC will be classified with statistically significant performances ranging between 70% and 80%[7] and (2) DL will perform better than traditional shallow approaches.[26]

Methods

Subjects

Participants were recruited as part as 5 independent studies carried out in multiple sites, all of which have been previously published:

- Site 1: Chengdu, China[29]
- Site 2: London, England (Genetic and Psychosis study[30])
- Sites 3 and 4: Santander A and B, Spain (Programa Asistencial Fases Iniciales de Psicosis (First Episode Psychosis Clinical Program) study[31])
- Site 5: Utrecht, The Netherlands (Genetic Risk and Outcome of Psychosis study[32])

All patients were experiencing their first psychotic episode, defined as the first manifestation of psychotic symptoms meeting criteria for a psychotic disorder, as specified by the DSM-IV[33] or ICD-10[34]. The demographic and clinical characteristics, including duration of illness, are reported in table 1. For information on recruitment criteria, see supplementary material.

MRI Data Acquisition and Preprocessing

High-resolution three-dimensional T1-weighted images were acquired independently at each site (supplementary table 2). From each image, 3 types of data features were extracted (see supplementary material):

- Voxel-based gray matter volume (GMV): whole-brain voxel-wise estimate of the local density of gray matter (GM) in a given voxel region[35]
- Voxel-based cortical thickness (VBCT): cortical thickness maps in which each voxel in the GM is assigned a thickness value[36,37]

Table 1. Demographic and Clinical Characteristics for FEP and HC for Each Site

		Chengdu, China (N = 222)		London, England (N = 142)		Santander A, Spain (N = 220)		Santander B, Spain (N = 210)		Utrecht, The Netherlands (N = 162)	
		HC	FEP	HC	FEP	HC	FEP	HC	FEP	HC	FEP
n		111	111	71	71	110	110	70	140	81	81
Gender (%)	M	51 (46)	51 (46)	36 (51)	36 (51)	68 (62)	68 (62)	45 (64)	90 (64)	64 (79)	64 (79)
	F	61 (54)	61 (54)	35 (49)	35 (49)	42 (38)	42 (38)	25 (46)	50 (46)	17 (21)	17 (21)
		$\chi 2$ = ns		$\chi 2$ = ns		$\chi 2$ = ns		$\chi 2$ = ns		$\chi 2$ = ns	
Age M (SD)		27.2 (7.3)	25.7 (8.1)	26.8 (7.1)	26.4 (6.2)	29.7 (7.8)	28.5 (8.6)	27.3 (7.5)	28.3 (7.6)	26.9 (8.0)	25.2 (5.9)
		t = ns		*t* = ns		*t* = ns		*t* = ns		*t* = ns	
TIV (L) M (SD)		1.5 (0.1)	1.5 (0.2)	1.5 (0.2)	1.5 (0.2)	1.5 (0.1)	1.4 (0.2)	1.5 (0.1)	1.5 (0.1)	1.6 (0.1)	1.5 (0.2)
		t = ns		*t* = ns		*t* = ns		*t* = ns		*t* = ns	
Positive symptoms M (SD)		—	24.6 (6.6)[a]	—	13.9 (5.5)[a]	—	14.7 (4.6)[b]	—	14.4 (4.1)[b]	—	15.9 (6.3)[a]
Negative symptoms M (SD)		—	18.2 (7.7)[a]	—	16.0 (6.0)[a]	—	6.3 (4.6)[c]	—	6.1 (5.0)[c]	—	16.2 (6.9)[a]
Duration of illness (years) Med (IQR)		—	0.3 (1.1)	—	1.1 (0.3)	—	0.3 (0.7)	—	0.3 (0.9)	—	0.6 (1.0)

Note: TIV, total intracranial volume; L, liters; M, male; F, female; FEP, first episode psychosis; HC, healthy controls, SD, standard deviation; Med, median; IQR, interquartile range.
[a]PANSS: Positive and Negative Symptoms Scale.
[b]SAPS: Scale for the Assessment of Negative Symptoms.
[c]SANS: Scale for the Assessment of Negative Symptoms.
ns: $P > .05$

- Surfaced-based regional volumes and cortical thickness: volume and thickness of predefined cortical and subcortical regions extracted with FreeSurfer[38]

Statistical Analysis

Demographic and Clinical Variables. Differences in age, gender, and total intracranial volume between FEP and HCs were examined using an independent-samples *t*-test and chi-square test, as implemented in the Statistical Package for the Social Sciences 24.0 (SPSS 24.0).

Group-Level Comparisons. For completeness, a standard group-level analysis was also carried out for each site and type of feature set separately. See supplementary material sections 1.4.1. and 2.1 for methods and results, respectively.

Multivariate Pattern Recognition Analysis. *Dimensionality Reduction: Principal Component Analysis* Principal component analysis (PCA) was used to reduce the number of voxels of the GMV and VBCT maps (see supplementary material).

Classifiers Four methods were used for classification: k-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM) and deep neural networks (DNN) (see supplementary material). These methods were chosen based on their increasing order of complexity (KNN is a straightforward algorithm, whereas DL can be more powerful at the expense of transparency), popularity (SVM and LR are among the most ML techniques used in previous studies), and novelty (DL has yielded promising results in psychiatric neuroimaging but is yet to be applied to FEP) (figure 1).

KNN: non-parametric method that uses the distance between data points to make new predictions by assigning unseen data to the same class to which the closest data points belong to[39].

LR: regression model applied to one dependent categorical variable implemented via elastic net, a regularized regression that combines the regularizations L1 and L2 penalties of Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression, respectively, to avoid overfitting.[40]

SVM: method that estimates a hyperplane with an optimum margin that best separates two classes, determined by the maximum distance from any data point. Once defined, this hyperplane is used to classify unseen data.[41,42]

DNN: multi-layered fully connected networks in which higher-level features are learned as a nonlinear combination of lower-level features, allowing the extraction of complex and abstract patterns.[43]

Model Training and Testing

Within-site classification. All models were assessed through a nested 10-fold stratified CV framework (figure 2) to ensure that the data for hyperparameter tuning and

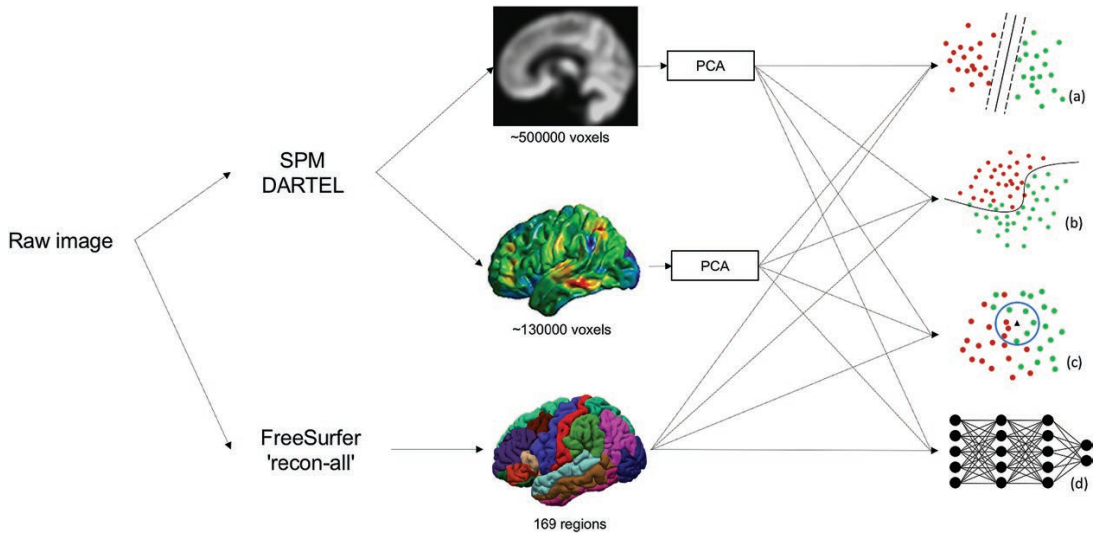


Fig. 1. Three features were extracted from each image: GMV, VBCT, and FreeSurfer surface-based regional volumes and cortical thickness. The dimensionality of GMV and VBCT was reduced through PCA. The resulting features were analyzed with four classifiers: (a) SVM, (b) LR, (c) KNN and (d) DNN. GMV, gray matter volume; VBCT, voxel-based cortical thickness; PCA, principal component analysis; SVM, support vector machine; KNN, k-nearest neighbors; LR, logistic regression; DNN, deep neural network.

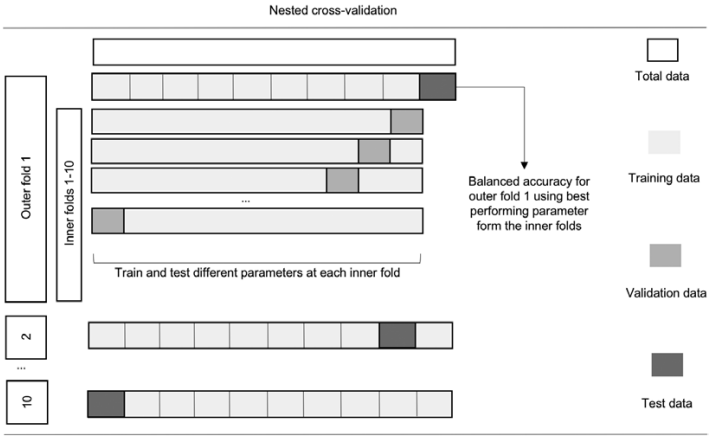


Fig. 2. Schematic representation of nested CV. Nested CV involves a secondary inner CV loop using the training data from the primary outer CV split, where different sets of hyperparameters are tested (eg, different values for the C parameter for SVM). The best-performing hyperparameters among the 10 inner folds are then used to train a model in the whole training set defined by the outer loop. This model is then tested using the test set of the outer loop. The final performance is estimated by averaging accuracies in the test set across all 10 outer folds. CV, cross-validation; SVM, support vector machine.

the data to test the algorithm were strictly independent. A 10-fold CV was chosen as a trade-off between bias, variance, and the demanding computational resources required to run DNN.

Cross-site classification. The best site-level model was further tested in each one of the remaining independent samples. All 10 instances trained during the CV were used to classify the participants from all the remaining sites separately. The resulting ensemble of models predicted the class of each participant using the soft voting method, where the class label was defined by the average of the 10 predicted probabilities.

Performance Measures Balanced accuracy, sensitivity, and specificity were chosen as the performance metrics. Statistical significance of the balanced accuracy was determined by permutation testing with 1000 permutations (see supplementary material).

Effect of Antipsychotic Medication and Psychotic Symptoms To examine whether antipsychotic medication or psychotic symptoms contributed to the classifiers' performance, chlorpromazine equivalents and positive and negative psychotic symptoms were regressed against the predicted labels using an logistic regression (see supplementary material for details).

Results

Sociodemographic and Clinical Parameters

No statistically significant differences were identified between patients and controls for age, gender, or total GMV at each site (table 1).

Single-Subject Classification

Can We Detect FEP at the Individual Level? Balanced accuracy, sensitivity, specificity, and statistical significance for each feature set of interest and site are presented in table 2 (for a visual display of the accuracies and standard deviations see supplementary figure 3 in the supplementary material). Overall, results were poor to modest across all types of feature sets and sites, although the site with the smallest sample size (site 2) showed the lowest performance consistently across all feature sets. Overall, regression analyses examining the effect of antipsychotic medication and psychotic symptoms on the performance of each classifier did not show a significant effect (see supplementary material).

What Are the Most Effective Type of Feature Set? There was no clear effect of type of feature set across sites. However, it can be seen that surface-based regional data tended to yield higher accuracies, especially when analyzed with DNN.

Can We Generalize the Results From One Site to the Others? The best performances were achieved by two DNN models at sites 1 and 3 using regional volumes and cortical thickness, with 70.5% and 70.2%, respectively. However, both models generalized poorly when tested on the remaining sites: specifically, the DNN model from site 1 achieved accuracies (sensitivity/specificity) of 52.1% (56.3%/47.9%), 61.1% (70.0%/52.7%), 52.1% (65.7%/38.6%), and 50.0% (48.3%/51.7%) when applied to sites 2 through 5, respectively; whereas the DNN model from site 3 achieved accuracies of 52.2% (96.5%/8.4%), 49.2% (83.5%/33.4%), 55.1%

Table 2. Accuracies (Sensitivity/Specificity) for Each Feature Set and Algorithm Across All Sites Using Nested 10-fold Stratified Cross-Validation. The Classifier Yielding the Best Balanced Accuracy Is Highlighted in Bold for Each Site

		Regional volumes and cortical thickness	GMV	VBCT
Site 1	KNN	60.7** (74.3/47.1)	**60.7** (49.5/71.9)**	62.1** (72.1/52.1)
Chengdu, China	LR	61.9** (64.9/58.9)	60.1** (62.9/58.6)	**67.2** (65.8/68.5)**
	SVM	61.3** (66.4/56.2)	**60.7** (63.0/58.5)**	52.7* (24.6/97.3)
	DNN	**70.5** (72.2/68.8)**	57.7** (59.5/56.0)	66.4** (63.9/68.3)
Site 2	KNN	56.7 (50.9/62.5)	43.9 (33.6/54.3)	53.5 (38.4/68.6)
London, England	LR	51.6 (45.0/58.2)	51.9 (53.8/50.0)	**61.6** (63.2/60.0)**
	SVM	45.9 (49.3/42.5)	**53.9 (53.4/54.3)**	51.0 (96.3/5.7)
	DNN	**58.8* (49.5/68.0)**	40.8 (47.4/34.3)	53.4 (52.4/55.3)
Site 3	KNN	59.6** (45.5/73.6)	50.5 (31.8/69.1)	58.0* (50.0/66.4)
Santander A, Spain	LR	58.6* (58.2/59.1)	63.2** (63.6/62.7)	59.1* (58.2/60.0)
	SVM	60.5** (61.8/59.1)	**65.9** (68.2/63.6)**	51.8* (90.9/12.7)
	DNN	**70.2** (70.0/70.4)**	50.2 (52.7/63.6)	**59.6 (60.0/59.1)**
Site 4	KNN	56.6* (91.8/21.4)	58.9** (70.7/47.1)	59.5* (67.7/51.1)
Santander B, Spain	LR	54.8 (73.9/35.7)	59.6** (57.8/61.4)	**62.6** (56.8/62.4)**
	SVM	56.0 (65.0/47.1)	57.4* (71.9/42.9)	58.4* (71.9/52.9)
	DNN	**62.0** (76.8/47.1)**	**59.3* (81.4/37.1)**	58.8** (62.4/53.1)
Site 5	KNN	52.7 (53.6/51.8)	54.5 (33.8/75.3)	52.2 (36.5/67.9)
Utrecht, The Netherlands	LR	58.5* (61.7/55.4)	61.3** (56.8/65.7)	**60.5** (60.6/60.4)**
	SVM	**60.7** (59.7/61.7)**	**62.4** (63.1/61.8)**	56.3 (51.2/61.4)
	DNN	54.9 (59.2/51.8)	58.0** (58.1/57.9)	60.1** (56.1/64.2)

Note: SVM, support vector machine; LR, logistic regression, KNN, k-nearest neighbors; DNN, deep neural network; GMV, voxel-based gray matter volume; VBCT, voxel-based cortical thickness.
*$P < .05$; **$P < .01$.

(70.1%/40.0%), and 51.0% (67.5%/34.6%) when applied to sites 1, 2, 4 and 5, respectively. To examine the possibility that poor generalizability was due to site differences, the same DNN model was applied to the total data with the 5 sites added as additional features. Features weights were then investigated to determine the importance of site. Results showed that out of the 174 features, the weights for site 1, 2, 3, 4, and 5 ranked 110, 150, 108, 71, and 112, respectively.

Discussion

In the last few years, there has been increasing interest in the translational potential of ML approaches in psychosis. As the field matures, there is emerging skepticism about replicability and generalizability, which has led to recent calls for greater caution in the interpretation of the findings.[9,17,22,23,25] This study aimed to elucidate the extent to which the application of ML to neuroanatomical data allows detection of individuals at the early stages of psychosis when the effects of illness chronicity and antipsychotic medication are minimal. To overcome the limitations of the existing literature, we used 5 independent datasets and put in place a series of methodological precautions to avoid overoptimistic results. Contrary to expectation, the performances of all methodological approaches tested were poor to modest across all sites. Later we discuss some of the main aspects that emerge from our investigation, including sample size, full independence of training and test data, cross-site generalizability, and testing multiple pipelines. We conclude the discussion by considering possible future directions.

Sample Size, Homogeneity, and Publication Bias

A possible explanation for why our accuracies are lower than those reported in the existing literature is that some of the previous studies may have reported overoptimistic results due to the use of fairly small sample sizes. To illustrate this possibility, we tested for an association between sample size and classification accuracy across studies using ML and structural MRI (sMRI) in the existing literature (see supplementary material). Unsurprisingly, we found a moderate negative association for studies that examined established schizophrenia ($r = -.41$) and FEP ($r = -.59$; after excluding Xiao et al,[20] which was a clear outlier; figure 3A). This is consistent with the notion that some of the previous studies may have reported overoptimistic accuracies due to the use of inadequate sample size.

There are at least two possible ways in which inadequate sample size can lead to an inflated estimation of the accuracy of an algorithm, including sample homogeneity and publication bias.[22,25] First, smaller samples tend to be more homogeneous, making it easier for an algorithm to learn shared abnormalities in patients relative to controls and resulting in higher accuracies. In contrast, larger samples tend to be more heterogeneous due to the loosening of inclusion criteria; in this case, it may be more challenging to find a shared pattern of abnormalities resulting in lower performances. This inverse relationship between

sample size and accuracy was not observed in our investigation; however, this might be explained by the fact that there was not sufficient variability in sample size across our five datasets. Second, smaller samples tend to be unstable and thus yield underestimated as well as overestimated accuracies.[21,44] This may, in turn, lead to publication bias, with overestimated accuracies being more likely to be published. In their meta-analysis of ML studies of schizophrenia, Kambeitz et al[7] reported that no publication bias was evident when all studies—including sMRI, functional magnetic resonance imaging, and DTI—were examined together. To test for publication bias in sMRI studies, we repeated the same statistical analysis focusing on this modality (see supplementary material). This revealed a statistically significant asymmetry in the funnel plot of published studies, indicating the presence of publication bias (figure 3B). This is in line with emerging concerns about possible overrepresentation of inflated performances in the literature.[17,22,23,25]

Full Independence of Training and Testing Set Data

Following recent recommendations on how to overcome methodological issues that may have led to initial inflated results,[9,23,25] we adopted two important methodological precautions. First, the use of simple CV, in which the same test data are used to both tune model hyperparameters and evaluate its performance, has been criticized as it almost certainly leads to inflated performances.[45,46] In the present investigation, algorithms were trained and tested via nested CV. This ensured that the test set remained fully independent from the training set, with only the latter being used to optimize model parameters. Second, implementing feature selection in a 2-step approach, where, eg, univariate tests (eg, *t*-test) are applied in the whole sample and only the statistically significant features are used for classification, is likely to result in overoptimistic performances as features are chosen based their performance on data that should be completely independent for testing the classifier. In the present investigation, therefore, transformations to the data, such as feature selection, were implemented within the CV framework, ie, parameters were derived from the training data only and subsequently applied to the test set. The adoption of these methodological precautions, aimed at ensuring full independence between training and test data, might explain the fact that accuracies in the present investigation were lower than expected.

Cross-Site Generalizability

The use of independent samples to develop and validate an algorithm is a critical requirement if the ultimate aim is to develop flexible ML-based tools that could be used in a clinical setting.[23,25] However, only a minority of studies have attempted to do this, eg,[22,47,48], and most of them have reported considerably lower performances in the independent sample. In the present investigation, the highest accuracies—obtained using specific combinations of dataset, type of feature set and algorithm—were 70% (in sites 1 and 3 with surface-based regional features and DNN); this performance would appear to be in line with previous similar studies. However, selectively reporting

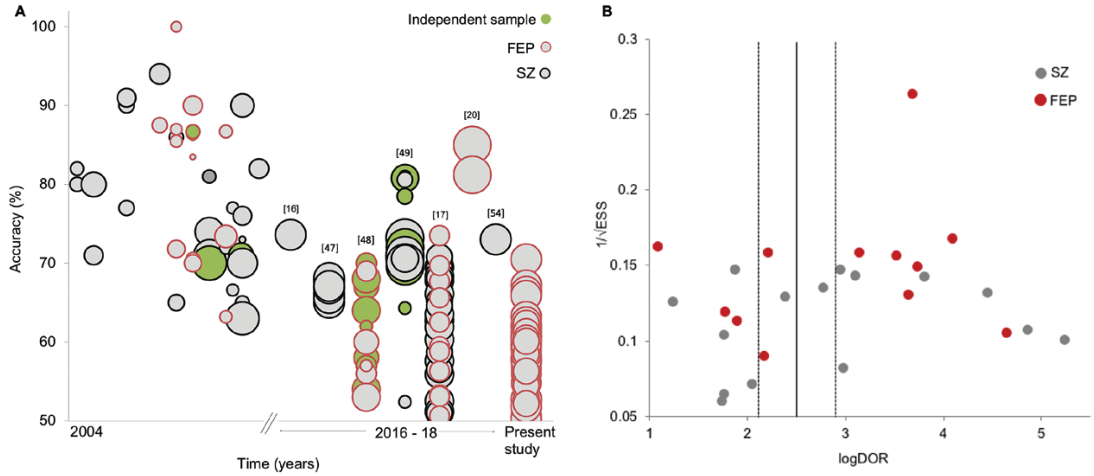


Fig. 3. (**A**) Accuracy of diagnostic sMRI ML studies over time and sample size (circle increases with sample size). From the first study until 2015, the vast majority of studies reported accuracies ranging between 70% and 100%; from 2016, however, performances have dropped overall with accuracies ranging between chance-level and 85%. (**B**) Funnel plot for sMRI studies in schizophrenia and FEP showing the distribution of individual studies according to their sample size (1/√ESS) and effect size (log diagnostic odds ratio). The plot revealed statistically significant asymmetric distribution around the main effect of sMRI studies ($P = .013$), indicating a bias favoring higher effect sizes. sMRI, structural MRI; ML, machine learning; FEP, first episode psychosis.

these accuracies from our wider set of results would have portrayed a distorted picture of the potential of ML to detect the initial stages of psychosis at the individual level.[24] This is especially true since after testing these two models in independent datasets, their performance did not hold up, indicating low cross-site generalizability. Such low cross-site generalizability could be due to site-related differences in scanning parameters, cultural interpretation of diagnostic criteria, and ethnicity; therefore, it might be possible to achieve higher cross-site generalizability by combining samples that are homogenous with respect to these variables. Nevertheless, our current results indicate that algorithms developed using data from a specific centre do not perform well when applied to data from other centers, and thus have limited clinical applicability.

Testing Multiple Pipelines

Because existing studies tend to differ with respect to several methodological aspects, at present, it is difficult to say which pipeline is optimal for detecting FEP.[47] Multi-pipeline studies have therefore been proposed as a useful way to disentangle what aspects works best.[23] Importantly, this approach may also help build more generalizable models, as the development of a bespoke, and possibility overfitted, pipeline to a local sample is less likely to occur. Consistent with this, Salvador et al[47] tested the performance of a range of ML approaches in different types anatomical features extracted from patients with schizophrenia and controls, and reported lower accuracies (66%–68%) compared to previous similar studies using a single pipeline. Winterburn et al[17] also used multiple pipelines in FEP and reported poor to modest accuracies, ranging from 51% to 73%. Taken collectively, evidence from these studies, including our own, suggest that when features are not manually carved to fit one algorithm applied to one specific small dataset, performance tends to drop. This can be seen in figure 3A where two generations of studies emerge: initially, there were mostly small single-site, single-feature, and single-algorithm high-performance studies; more recently the use of (1) larger samples,[16,47,20,54] (2) multicentre studies,[48,49] (3) assessment of different algorithms and/or features in one/several site(s),[17,47] or (4) independent sample testing[48,49] are reshaping the original, and possibly overinflated, enthusiasm with more realistic performances.

What Next for ML-sMRI Studies of Psychiatric Disease?

Unlike group-level analysis, where larger samples lead to increased chance of detecting a statistically significant result (even with a small effect size), in ML larger samples do not necessarily equate to better results; instead, these tend to lead to lower accuracies due to increased heterogeneity.[22,28] Despite this challenge, larger samples are likely to be more representative of the illness, less likely to overfit and thus carry more translational potential. Future ML studies will have to address this issue to overcome the increasingly apparent bottleneck in the performance that is arising with larger sample sizes (figure 3A). A possible way of doing so could be to use normative models, where an individual is mapped against a normative model that should encompass the heterogeneity characteristic of the normal population. Here, illness is considered an extreme case within a normal range, which is likely to be a more ecologically valid approach than the traditional case–control paradigm.[50,51]

Greater methodological standardization based on "good-practice recommendations" could also help disentangle the current conflicting evidence. For example, guidelines for minimum sample size such as the threshold ($n > 130$) proposed by Nieuwenhuis et al[21] are a good start. The need for independent sample testing has also been widely acknowledged as an essential step toward generalizability[23,25]; however, even the most recent studies do not always perform this. Moving forward, this type of generalizability test is likely to become a gold standard for ML diagnostic studies. More transparency in the implementation of ML is also needed. Several studies do not provide enough information about how the algorithm was trained and tested.[23,28,52] This hinders a thorough assessment of the validity of the study as well as its replicability. Finally, it should be noted that, even if sMRI was able to distinguish between patients with FEP and disease-free individuals with high levels of accuracy, this would be of limited clinical utility. This is because, from a clinical translation perspective, the real challenge is not to distinguish between patients and disease-free individuals, but to develop biological tests that could be used to choose between alternative diagnoses and optimize treatment.[52]

Conclusion

The present investigation attempted to overcome the limitations of the existing literature using a number of strategies. First, we studied patients with FEP in which the effects of antipsychotic medication and illness chronicity are likely to be minimal. Second, the sample size of each of our 5 datasets was greater than the recommended threshold for achieving a stable performance in ML–sMRI studies.[21] Third, critical methodological precautions (eg, nested CV and appropriate use of feature selection) were adopted to ensure an unbiased assessment of performance. Fourth, we systematically assessed the performance of a range of algorithms and features across several datasets, thereby minimizing the possibility of developing a bespoke and likely overfitted model to a single site. Fifth, we assessed the cross-site generalizability of the best models at the single-site level. Our

findings suggest that the use of ML and sMRI allows detection of FEP at the individual level with relatively modest accuracies—lower than what was expected based on previous studies and much lower than what would be required for clinical translation. We speculate that some of the previous results may have been over-optimistic due to a combination of small sample sizes, less-than-rigorous methodologies, and possible publication bias and argue that the current evidence for the diagnostic value of ML and structural neuroimaging should be reconsidered toward a more cautious interpretation.

Over the past few years, the number of ML studies in psychosis has been increasing rapidly.[52] As larger samples and more powerful computational resources become available, this momentum is likely to continue to grow over the coming years.[53] Therefore, it is important for the research community to be aware of the challenges and limitations of applying ML to psychosis such as the several potential "distortion" of the findings along the ML pipeline, as discussed in a recent review.[52] In light of these challenges and limitations, the extent to which the application of ML in psychosis will lead to a more valid construct of the illness remains an open question. We encourage researchers to continue pursuing the integration of ML and neuroimaging, while exercising caution to avoid inflated results and ultimately a distorted view of the potential of this approach in psychiatric neuroimaging.

Supplementary Material

Supplementary data are available at *Schizophrenia Bulletin* online.

Funding

This work was supported by the European Commission (PSYSCAN—Translating neuroimaging findings from research into clinical practice; 603196 to P.M.); International Cooperation and Exchange of the National Natural Science Foundation of China (81220108013 to Q.G. and A.M.); Wellcome Trust's Innovator Award (208519/Z/17/Z to A.M.); Foundation for Science and Technology (SFRH/BD/103907/2014 to S.V.), and São Paulo Research Foundation (FAPESP) (Brazil; 2013/05168-7 to W.H.L.P.). The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Chan RC, Di X, McAlonan GM, Gong QY. Brain anatomical abnormalities in high-risk individuals, first-episode, and chronic schizophrenia: an activation likelihood estimation meta-analysis of illness progression. *Schizophr Bull.* 2011;37(1):177–188.
2. Fusar-Poli P, Borgwardt S, Crescini A, et al. Neuroanatomy of vulnerability to psychosis: a voxel-based meta-analysis. *Neurosci Biobehav Rev.* 2011;35(5):1175–1185.
3. Torres US, Duran FL, Schaufelberger MS, et al. Patterns of regional gray matter loss at different stages of schizophrenia: a multisite, cross-sectional VBM study in first-episode and chronic illness. *Neuroimage Clin.* 2016;12:1–15.
4. Smieskova R, Fusar-Poli P, Allen P, et al. Neuroimaging predictors of transition to psychosis–a systematic review and meta-analysis. *Neurosci Biobehav Rev.* 2010;34(8):1207–1222.
5. Vita A, De Peri L, Deste G, Sacchetti E. Progressive loss of cortical gray matter in schizophrenia: a meta-analysis and meta-regression of longitudinal MRI studies. *Transl Psychiatry.* 2012;2(11):e190.
6. Davatzikos C, Shen D, Gur RC, et al. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch Gen Psychiatry.* 2005;62(11):1218–1227.
7. Kambeitz J, Kambeitz-Ilankovic L, Leucht S, et al. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology.* 2015;40(7):1742–1751.
8. Zarogianni E, Moorhead TW, Lawrie SM. Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. *Neuroimage Clin.* 2013;3:279–289.
9. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev.* 2015;57:328–349.
10. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev.* 2012;36(4):1140–1152.
11. Navari S, Dazzan P. Do antipsychotic drugs affect brain structure? A systematic and critical review of MRI findings. *Psychol Med.* 2009;39(11):1763–1777.
12. Vita A, De Peri L, Deste G, Barlati S, Sacchetti E. The effect of antipsychotic treatment on cortical gray matter changes in schizophrenia: does the class matter? a meta-analysis and meta-regression of longitudinal magnetic resonance imaging studies. *Biol Psychiatry.* 2015;78(6):403–412.
13. Bora E, Fornito A, Radua J, et al. Neuroanatomical abnormalities in schizophrenia: a multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophr Res.* 2011;127(1-3):46–57.
14. van Erp TGM, Hibar DP, Rasmussen JM, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol Psychiatry.* 2016;21(4):547–553.
15. van Erp TGM, Walton E, Hibar DP, et al.; Karolinska Schizophrenia Project. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biol Psychiatry.* 2018;84(9):644–654.
16. Pinaya WH, Gadelha A, Doyle OM, et al. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep.* 2016;6(1):38897.
17. Winterburn JL, Voineskos AN, Devenyi GA, et al. Can we accurately classify schizophrenia patients from healthy

controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophr Res.* 2017 Dec 20. pii: S0920-9964(17)30736-3. doi:10.1016/j.schres.2017.11.038. [Epub ahead of print]

18. Pettersson-Yeo W, Benetti S, Marquand AF, et al. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol Med.* 2013;43(12):2547–2562.
19. Borgwardt S, Koutsouleris N, Aston J, et al. Distinguishing prodromal from first-episode psychosis using neuroanatomical single-subject pattern recognition. *Schizophr Bull.* 2013;39(5):1105–1114.
20. Xiao Y, Yan Z, Zhao Y, et al. Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophr Res.* 2017 Dec 2. pii: S0920-9964(17)30735-1. doi: 10.1016/j.schres.2017.11.037. [Epub ahead of print]
21. Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage.* 2012;61(3):606–612.
22. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front Psychiatry.* 2016;7:50.
23. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage.* 2017;145:137–165.
24. Janssen RJ, Mourão-Miranda J, Schnack HG. Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biol Psychiatry Cogn Neurosci Neuroimaging.* 2018;3(9):798–808.
25. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci.* 2017;20(3):365–377.
26. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev.* 2017;74:58–75.
27. Plis SM, Hjelm DR, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study. *Front Neurosci.* 2014;8:229.
28. Schnack HG. Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr Res.* 2017 Oct 24. pii: S0920-9964(17)30649-7. doi:10.1016/j.schres.2017.10.023. [Epub ahead of print]
29. Gong Q, Dazzan P, Scarpazza C, et al. A neuroanatomical signature for schizophrenia across different ethnic groups. *Schizophr Bull.* 2015;41(6):1266–1275.
30. Di Forti M, Morgan C, Dazzan P, et al. High-potency cannabis and the risk of psychosis. *Br J Psychiatry.* 2009;195(06):488–491.
31. Pelayo-Terán JM, Pérez-Iglesias R, Ramírez-Bonilla M, et al. Epidemiological factors associated with treated incidence of first-episode non-affective psychosis in Cantabria: insights from the Clinical Programme on Early Phases of Psychosis. *Early Interv Psychiatry.* 2008;2(3):178–187.
32. Korver N, Quee PJ, Boos HB, Simons CJ, de Haan L; GROUP investigators. Genetic Risk and Outcome of Psychosis (GROUP), a multi-site longitudinal cohort study focused on gene-environment interaction: objectives, sample characteristics, recruitment and assessment methods. *Int J Methods Psychiatr Res.* 2012;21(3):205–221.
33. APA. *Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV-TR)*. Washington, DC: American Psychiatric Association; 2000.
34. Organization World Health. *International Classification of Diseases, Tenth Revision*. Geneva, Switzerland: World Health Organization; 1992.
35. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage.* 2007;38(1):95–113.
36. Hutton C, De Vita E, Ashburner J, Deichmann R, Turner R. Voxel-based cortical thickness measurements in MRI. *Neuroimage.* 2008;40(4):1701–1710.
37. Hutton C, Draganski B, Ashburner J, Weiskopf N. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *Neuroimage.* 2009;48(2):371–380.
38. Fischl B. FreeSurfer. *Neuroimage.* 2012;62(2):774–781.
39. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat.* 1992;46(3):175–185.
40. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol).* 2005;67(2):301–320.
41. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage.* 2009;45(1):S199–S209.
42. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
43. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–444.
44. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage.* 2018;180:68–77.
45. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv.* 2010;4:40–79.
46. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7(1):91.
47. Salvador R, Radua J, Canales-Rodríguez EJ, et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS One.* 2017;12(4):e0175683.
48. Dluhoš P, Schwarz D, Cahn W, et al. Multi-center machine learning in imaging psychiatry: a meta-model approach. *Neuroimage.* 2017;155:10–24.
49. Rozycki M, Satterthwaite TD, Koutsouleris N, et al. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophr Bull.* 2018;44(5):1035–1044.
50. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry.* 2016;80(7):552–561.
51. Sato JR, Rondina JM, Mourão-Miranda J. Measuring abnormal brains: building normative rules in neuroimaging using one-class support vector machines. *Front Neurosci.* 2012;6:178.
52. Tandon N, Tandon R. Will machine learning enable us to finally cut the gordian knot of schizophrenia. *Schizophr Bull.* 2018;44(5):939–941.
53. Bzdok D, Yeo BTT. Inference in the age of big data: future perspectives on neuroscience. *Neuroimage.* 2017;155:549–564.
54. de Moura AM, Pinaya WHL, Gadelha A, et al. Investigating brain structural patterns in first episode psychosis and schizophrenia using MRI and a machine learning approach. *Psychiatry Res Neuroimaging.* 2018;275:14–20.

Supplementary material

Contents

eTables

1. eMethods

1.1. Participants

1.1.1. Recruitment procedure and criteria

Site1: Chengdu, China

First episode patients were recruited from the West China Hospital of Sichuan University in Chengdu (China), as part of a wider study of psychiatric disorders in China. Diagnosis and duration of illness were determined by the consensus of two clinical psychiatrists using the Structured Interview for the DSM-IV Axis I Disorder (SCID)[1]. At the time of scanning, all patients were medication-naïve. Healthy controls were recruited by poster advertisement and screened using the SCID-I to confirm the lifetime absence of psychiatric disorders, as well as interviewed and subsequently excluded if they had any known history of psychiatric illness in first-degree relatives. Participants were excluded if they met any of the following criteria: (i) history of drug or alcohol abuse, (ii) pregnancy, and (iii) any physical illness such as hepatitis, cardiovascular disease, or neurological disorder, as assessed by interview and review of medical records.

Site 2: London, England

Participants were recruited from the South London and Maudsley Foundation Trust and scanned at the Institute of Psychiatry, Psychology and Neuroscience in London (England). Diagnosis of schizophrenia was formulated by an experienced psychiatrist using the ICD-10 criteria. Healthy controls were recruited through local advertisement from the same geographical areas as patients. A screening tool (Psychosis Screening Questionnaire[2]) was used to exclude the presence of psychotic symptomatology or a history of psychotic illness. Additional exclusion criteria for all participants included learning disabilities (based as an IQ < 70), current or past neurological illness, brain injury with loss of consciousness for more than 1 hour and suspected or confirmed pregnancy.

Sites 3 and 4: Santander A and B, Spain

Data from Santander A and Santander B were acquired as part of the same large prospective longitudinal study on first episode psychosis in the region of Cantabria, although with two different scanners (eTable 2). Individuals with FEP were recruited from both inpatient units and community services throughout the entire region. Patients were included if they met the following criteria: 1) age 15–60 years; 2) DSM-IV criteria for a principal diagnosis of schizophrenia, schizophreniform disorder, schizoaffective disorder, brief reactive psychosis, or not otherwise specified psychosis; and 3) no prior treatment with antipsychotic medication or, if previously treated, a total lifetime of adequate antipsychotic treatment of less than 6 weeks. Patients with DSM-IV based diagnoses of mental retardation or substance dependence (except nicotine dependence) were excluded. Age and gender matched healthy controls were recruited from the community through advertisements and were screened for current or past history of psychiatric, mental retardation, neurological or general medical illness, including substance dependence and significant loss of consciousness, as determined by using an abbreviated version of the Comprehensive Assessment of Symptoms and History (CASH)[3]. Clinical records and family interview also confirmed the absence of psychosis in first-degree relatives.

Site 5: Utrecht, The Netherlands

Inpatients and outpatients were identified by clinicians working in regional psychosis departments or academic centres and were included if they met the following criteria: 1) age range of 16 to 50 years; 2) a diagnosis of non-affective psychotic disorder according to the DSM-IV; 3) good command of the Dutch language; and 4) able and willing to give written informed consent. Controls were selected through a system of random mailings to addresses in the catchment areas of the cases and were included if the following criteria were met: 1) age range of 16 and 50 years, 2) no lifetime psychotic disorder, 3) no first-degree family member with a lifetime psychotic disorder, 4) good command of the Dutch language, and 5) able and willing to give written informed consent.

1.1.2. Matching

To maximize the use of the data made available, matching was carried out by taking the group with smallest sample size (first episode psychosis (FEP) or healthy controls (HC)) and randomly selecting participants from the other group according to age (+/- 5 years) and gender. For all sites, the FEP:HC matching ratio was 1:1, except for site 4 where the ratio was 2:1.

eTable 1. Sample size of each dataset. We report the number of subjects available (top row), the number of subjects excluded after matching patients and controls for age and gender (middle row) and the final number of subjects included in the statistical analysis (bottom row).

	Site 1 Chengdu, China	Site 2 London, England	Site 3 Santander A, Spain	Site 4 Santander B, Spain	Site 5 Utrecht, The Netherlands
Available	330	204	257	223	225
Excluded	108	62	37	13	63
Final	222	142	220	210	162

1.2. MRI data acquisition

eTable 2. Image acquisition parameters for each site.

	Site 1	Site 2	Site 3	Site 4	Site 5
Field strength (T)	3	3	3	1.5	1.5
TR/TE (ms)	8.5/3.4	6.9/2.8	8.2/3.7	24/5	30/4.6
Slice thickness (mm)	1	1.2	1	1.5	1.2
Data matrix	512x512x156	256x256x166	256x256x160	256x256x124	256x256x170
Voxel size	0.47x0.47x1	1.02x1.02x1.2	0.94x0.94x1	1.02x1.02x1.5	1x1x1.2

1.3. MRI preprocessing

After checking all T1-weighted images for scanner artefacts and gross anatomical abnormalities, images were preprocessed to extract three types of anatomical features: voxel-based grey matter volume, voxel-based cortical thickness and surface-based volumes and cortical thickness.

1.3.1. Voxel-based maps

Two different voxel-based features were extracted: grey matter volume and cortical thickness. Common to both features, images were first reoriented along the anterior-posterior commissure line and set the anterior commissure as the origin of the spatial coordinates to assist the normalization algorithm. Reoriented images were then segmented into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) partitions as implemented in SPM12[4] (http://www.fil.ion.ucl.ac.uk/spm).

1.3.1.1. Grey matter volume

The segmentation tissue maps for each site were pre-processed separately using the Diffeomorphic Anatomical Registration using the Exponentiated Lie algebra (DARTEL)

toolbox[5]. This procedure warps the grey matter and white matter partitions into a new study-specific reference space representing an average of all the subjects included in the analysis, thus maximizing accuracy and sensitivity[6,7]. The warped grey matter partitions were then affine-transformed into MNI space. An additional modulation step was used to scale the grey matter probability values by the Jacobian determinants of the deformations, thereby ensuring that the total amount of grey matter in each voxel was conserved after registration[8]. Finally, the GM probability maps were smoothed using a standard 8mm FWHM Gaussian kernel.

1.3.1.2. Cortical thickness

A voxel-based Laplacian method[9], implemented as an SPM toolbox[10,11], was used to create a voxel-based cortical thickness (VBCT) map for each subject using the GM, WM and CSF partitions generated in the segmentation step. Briefly, the resulting VBCT maps contained cortical thickness (CT) values within voxels identified as grey matter and zeros outside the cortex. Each VBCT map was warped into the corresponding site-specific DARTEL reference space. The warped images were then normalized to MNI space and smoothed with a 6 mm Gaussian kernel. The same warps, modulation and smoothing were also applied to a binary mask created from each original VBCT map. Subsequently the warped, scaled and smoothed VBCT maps were divided by the corresponding warped, scaled, and smoothed mask.

1.3.2. Surface-based volume and cortical thickness

FreeSurfer 5.3 (http://surfer.nmr.mgh.harvard.edu)[12] was used to parcellate each participant's raw brain image into subcortical and cortical regions according to the Desikan-Killiany atlas[13] using the 'recon-all' command. FreeSurfer is a well-established automated procedure for imaging preprocessing and analysis which details have been extensively described elsewhere[14–16]. A total of 169 features were used, including 33 volumes of subcortical structures plus volume and thickness of 34 cortical regions per hemisphere (after removing white matter hypo-intensities, 5th ventricle, optic chiasm and bilateral vessels and choroid plexus).

1.4. eStatistical analysis

1.4.1. Group-level analysis

1.4.1.1. Grey matter volume and cortical thickness

Voxel-based morphometry (VBM) was used to calculate group-level differences in voxel-based grey matter volume and voxel-based cortical thickness between FEP and HC groups at each site. An independent-sample t-test was used with statistical inferences made at $p<0.05$ after family-wise error (FWE) correction for multiple comparisons and a minimum extent threshold of 5 voxels.

1.4.1.2. Surface-based regional volumes and cortical thickness

Surface-based regional volumes and cortical thickness were analysed with an independent-sample t-test as implemented in SPSS 24.0 using a statistical threshold of $p<0.05$ and additional Bonferroni correction for multiple comparisons.

All reported results (eResults, section 2.1) were obtained without covariates of no interest to ensure consistency between group- and individual-level statistical analyses. However, statistical analyses with age and gender as covariates were also carried out for completeness; this yielded identical results except for surface-based regional volumes and cortical thickness data (eTable 7-9).

1.4.2. Multivariate pattern recognition analysis

1.4.2.1. Dimensionality reduction: principal component analysis

Principal component analysis (PCA) is a well-established unsupervised method for feature reduction in neuroimaging. PCA reduces dimensionality by geometrically projecting the data into lower dimensions called principal components (PCs), with the aim of finding the best summary of the data using a limited number of PCs. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated features into a set of values of

uncorrelated features (PC). PCs are then ranked according to explained variance in descending order. A detailed description of PCA is given elsewhere[17,18]. In the present investigation, PCA was implemented within the CV framework; at each fold, dimensionality was reduced by 1) extracting the minimum number of principal components whilst retaining cumulative 90% of the variance from the data in the training set only, 2) projecting all grey matter/cortical thickness maps onto the resulting principal components and 3) using the resulting values for classification and 4) projecting the test data into the same components derived from the training set, and using the former for testing.

1.4.2.2. Feature scaling: Standardization

Standardization was performed by removing the mean and scaling to unit variance. This procedure was applied to each feature independently. Standardization is a common requirement for many ML methods, since algorithms might behave poorly if the individual features do not resemble normally distributed data. In addition, features with bigger scales might dominate the loss function of the training algorithms. To avoid “double dipping”, the statistics (mean and variance) were obtained using only the training set, and these same values were used in the standardization of test set.

1.4.2.3. Classifiers

K-nearest neighbour (KNN), logistic regression (LR) and support vector machine (SVM) were implemented using the Scikit-Learn library[19] (sklearn) for python 3.5. Deep neural network (DNN) was implemented using Tensorflow v.1.4[20] and Keras v.2.1[21] libraries. The random seed was kept the same for all models to ensure the reproducibility of the results. This approach guaranteed that the starting weights and train/test split at each fold of the CV would remain the same within and between algorithms for the same site.

1.4.2.3.1. K-nearest neighbours

K-nearest neighbours (KNN) is a non-parametric method based on multivariate pairwise distance measures between data points. Once presented with unseen data, it calculates the Euclidean distance between this new data point and each of the surrounding neighbours. Classification is done by assigning the unseen data to the same class as the majority of its neighbours[22]. The optimal number of neighbours was tuned via grid search by testing 10 possible odd values ranging from 3 to 21 in increments of 2.

1.4.2.3.2. Logistic regression

Logistic regression (LR) was implemented via elastic net, a regularized regression that combines the regularizations L1 and L2 penalties of LASSO (Least Absolute Shrinkage and Selection Operator) and ridge regression, respectively. While the ridge penalty retains all variables and minimizes the impact of irrelevant features, the LASSO penalty discards unimportant variables[23]. Grid search was used to find the optimal relative contribution of each penalty via tuning of the hyperparameter l1_ratio as defined by sklearn from eleven possible values between 0 and 1 with increments of 0.1.

1.4.2.3.3. Support vector machine

Support vector machine (SVM) is a supervised machine learning technique that maps the input data into a feature space using a set of similarity functions known as kernels. In this feature space, the model finds the optimal separating hyperplane by finding the largest margin of separation between the two classes within the training set. Once the hyperplane is determined, it can be used to predict the class of new unseen observations[24,25]. In this study, a linear kernel was chosen to contrast with the characteristic non-linear approach of DL. The soft margin (C) parameter, that controls the trade-off between having zero training errors and allowing misclassifications, was tuned from a possible range of values (2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}) using grid search, i.e. all possible values in a given range were tested.

1.4.2.3.4. Deep neural network

Given its flexible architecture, deep learning can be used to build a variety of different neural networks[26]. Here we employed a deep neural network, with the components resulting from the PCA (for the VBM and VBCT data) or the regional volumes and cortical thickness as inputs; this architecture was chosen as it allowed for automated and non-biased optimization of the hyperparameters, which in turn helps prevent overfitting. Deep neural networks are multi-layered fully-connected networks where higher-level features are learned as a non-linear combination of lower-level features, thus allowing the extraction of complex and abstract patterns from the data. Once the model learns these higher-level features, it can determine a separation surface to classify the different classes[26,27]. The performance of DNN models relies on the specification of several architectural and learning hyperparameters. To prevent bias, the number of layers, number of units, optimizer, learning rate, decay, activation function and epoch and were optimized using random search as implemented by sklearn. To decrease the chances of overfitting, two additional parameters were also included at each layer: i) L2 regularizer, which penalizes high weights[28] and ii) dropout, where randomly selected neurons are ignored during training[29]. Each layer was initialized via Glorot (also known as Xavier) initialization (normal distribution)[30]. In the output layer, the classification was performed by a softmax function. Training was carried out using a mini-batch with 8 training samples for VBM and VBCT, and 128 for surface-based volumes and cortical thickness. DNN models were optimized via random search due to the high number of parameters to test: at each fold, 500 different combinations of randomly selected values for each parameter were tested. eTable 3, shows all the possible values for each parameter.

eTable 3. Parameters for tuning for DNN

Parameter	Values
Number of layers	2, 3, 4, 5
Number of units	10, 20, 50, 75, 100, 150
Activation function	ReLU, Leaky ReLU
Learning rate	0.001, 0.005, 0.01, 0.1, 0.2
Learning rate decay	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}
Epochs	50, 100, 150
Optimizer	Stochastic gradient descent (SGD), Adam
Momentum	0.99, 0.9, 0.95
L2 coefficient	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}
Drop-out rate	0.2, 0.5, 0.7

1.4.2.4. Performance measures

Performance metrics were calculated according to the below formulas:

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Balanced accuracy = (Sensitivity + Specificity)/2

1.4.2.5. Significance testing

The balanced accuracy of each classifier was tested for significance using permutation testing, whereby subjects were randomly assigned to one of the classes (patients/control), so that the labels no longer match the data in any meaningful way, and the 10-fold CV cycle repeated 1000 times. This resulted in a distribution of accuracies reflecting the null hypothesis that the classifier did not exceed chance. The number of times the classifier's performance was greater

than or equal to the true accuracy was divided by 1000 to determine a p-value. A p-value lower than 0.05 was considered statically significant.

1.4.2.6. Effect of medication and psychotic symptoms

To examine whether anti-psychotic medication and psychotic symptoms contributed to the classifiers' performance, chlorpromazine equivalents and positive and negative psychotic symptoms were regressed against the predicted labels using a logistic regression as implemented by the Logit function from the statsmodel python library. Because all patients from site 1 were anti-psychotic naïve, the investigation of the effects of medication was limited to sites 2, 3, 4 and 5. The size of the effects of medication and psychotic symptoms was measured in terms of odds ratio (OR) and respective 95% confidence interval (CI). The statistical significance threshold was set to 0.05.

2. eResults

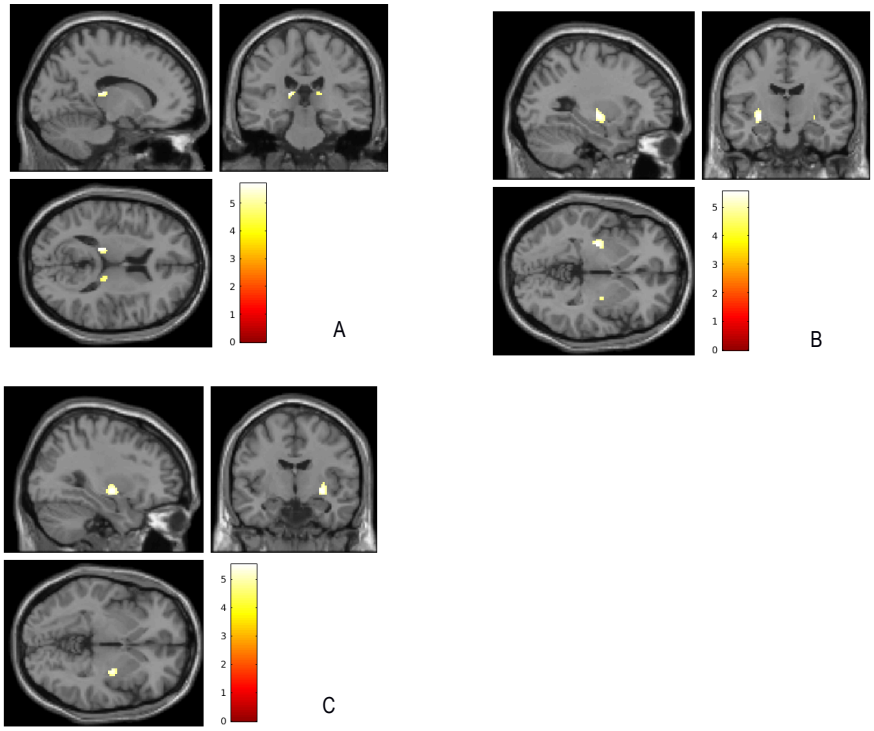
2.1. Group-level analyses

No significant GMV decreases in FEP relative HC were found at any site. In contrast, GMV increases were detected in the bilateral thalamus at site 3; in the left putamen and the right pallidum at site 4; and in the right putamen at site 5. No significant increased or decreased VBCT was observed in FEP compared to HC at any site, except for site 1 in which FEP showed increased thickness in the left fusiform gyrus and left superior frontal gyrus. Significant differences in surface-based regional volumes and cortical thickness between FEP and HC were found for sites 3 and 4. At site 3, patients showed smaller right hippocampus volume as well as a reduced thickness of the inferior parietal lobe; whereas at site 4 patients showed a significant cortical thinning in the left inferior temporal gyrus, pars opercularis and rostral middle frontal gyrus, as well as a larger 3rd ventricle. These results are presented in detail in eTables 4-7.

eTable 4. Group-level analysis: GMV.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 3				
Left thalamus	-16,-28,12	37	5.5	.006
Right thalamus	16,-24,14	17	4.7	.014
Site 4				
Left putamen	-30,-12,-4	118	5.3	.001
Right pallidum	20,-10,-4	17	4.6	.014
Site 5				
Right putamen	30,-8,-4	85	5.3	.001

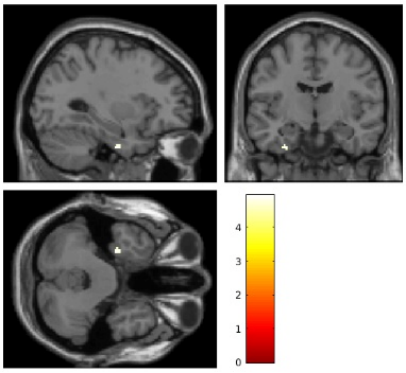
eFigure 1. Regions with increased GMV in FEP relative to controls in site 3 (A,) 4 (B) and 5 (C).



eTable 5. Group-level analysis: VBCT.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 1				
Left fusiform gyrus	-30,-10,-36	11	4.7	.012
Left superior frontal gyrus	-10,58,36	11	4.5	.012

eFigure 2. Cortical region with increased GMV in FEP relative to controls in site 1.



eTable 6. Group-level analysis: surface-based regional volumes and cortical thickness.

Region	*t*	*p*
FEP < HC		
Site 3		
Right hippocampus	4.3	<.001
Left inferior parietal (thickness)	3.9	<.001
Site 4		
Left inferior temporal gyrus (thickness)	3.6	<.001
Left pars opercularis (thickness)	4.0	<.001
Left rostral middle frontal gyrus (thickness)	4.8	<.001
FEP > HC		
Site 4		
Third ventricle	-4.1	<.001

eTable 7. Group-level analysis controlling for age and gender: GMV.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 3				
Left thalamus	-16,-28,12	40	5.6	.006
Right thalamus	16,-24,14	19	4.8	.014
Site 4				
Left putamen	-30,-14,-2	110	5.1	.001
Right pallidum	28,-10,-4	17	4.6	.014
Site 5				
Right putamen	30,-8,-6	70	5.2	.002

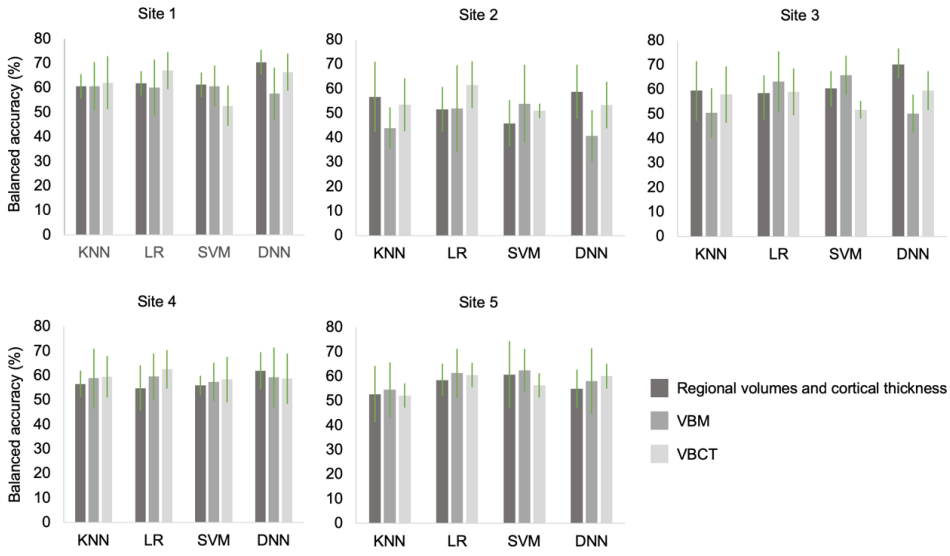
eTable 8. Group-level analysis controlling for age and gender: VBCT.

Region	Peak MNI Coordinates (x,y,z)	Cluster size (No. of Voxels)	z	*p*
FEP > HC				
Site 1				
Left fusiform gyrus	-30,-10,-36	16	5.0	.008
Left superior frontal gyrus	-10,58,36	11	4.6	.012

eTable 9. Group-level analysis controlling for age and gender: surface-based regional volumes and cortical thickness.

Region	*F*	*p*
FEP < HC		
Site 3		
Right hippocampus	22.4	<.001
Left inferior parietal gyrus (thickness)	20.7	<.001
Left precuneos (thickness)	15.3	<.001
Left superior frontal gyrus (thickness)	12.8	<.001
Left supramarginal gyrus (thickness)	17.2	<.001
Site 4		
Left parsopercularis (thickness)	15.6	<.001
Left rostral middle frontal gyrus (thickness)	21.9	<.001
Site 5		
Left hippocampus	15.7	<.001
FEP > HC		
Site 3		
Left lateral ventricle	14.8	<.001
Site 4		
Third ventricle	17.2	<.001

eFigure 3. Balanced accuracies and standard deviations of the different algorithms and feature sets for each site.



KNN: k-nearest neighbours; LR: logistic regression; SVM: support vector machines; DNN: deep neural network; VBM: voxel-based morphometry; VBCT: voxel-based cortical thickness.

eTable 10. Statistical significance for all classifiers.

		Surface-based regional volumes and cortical thickness	VBM	VBCT
Site 1 Chengdu China	KNN	**.003**	**.001**	**.002**
	LR	**.003**	**.003**	**.001**
	SVM	**.003**	**.004**	**.013**
	DNN	**.001**	**.008**	**.001**
Site 2 London England	KNN	.083	.891	.200
	LR	.381	.346	**.009**
	SVM	.757	.207	.450
	DNN	**.014**	.593	.265
Site 3 Santander A Spain	KNN	**.004**	.444	**.011**
	LR	**.021**	**.002**	**.013**
	SVM	**.005**	**.001**	**.036**
	DNN	**.001**	.448	**.010**
Site 4 Santander B Spain	KNN	**.041**	**.003**	**.028**
	LR	.129	**.012**	**.001**
	SVM	.081	**.032**	**.030**
	DNN	**.001**	**.014**	**.002**
Site 5 Utrecht The Netherlands	KNN	.237	.163	.262
	LR	**.033**	**.003**	**.007**
	SVM	**.007**	**.004**	.408
	DNN	.108	**.010**	**.008**

eTable 11. Odds ratio, confidence interval and p-value for the effects of anti-psychotic medication and psychotic symptoms on predicted labels.

		Surface-based regional volumes and cortical thickness			VBM			VBCT		
		Anti-psychotic medication	Positive symptoms	Negative symptoms	Anti-psychotic medication	Positive symptoms	Negative symptoms	Anti-psychotic medication	Positive symptoms	Negative symptoms
SITE 1	KNN	-	0.99 [0.97-1.09], .656	0.98 [0.97-1.09], .575	-	1.03 [0.97-1.09], .333	1.0 [0.94-1.04], .587	-	1.03 [0.97-1.1], .334	1.0 [0.94-1.05], .768
	LR	-	1.0 [0.97-1.09], .896	1.03 [0.97-1.09], .264	-	0.97 [0.91-1.03], .276	1.0 [0.97-1.07], .516	-	1.01 [0.95-1.07], .783	1.0 [0.99-1.1], .117
	SVM	-	1.0 [0.97-1.09], .896	1.03 [0.97-1.09], .264	-	0.99 [0.93-1.05], .754	1.0 [0.93-1.03], .477	-	0.96 [0.87-1.06], .464	1.0 [0.89-1.08], .676
	DL	-	0.97 [0.97-1.09], .378	1.03 [0.97-1.09], .351	-	0.99 [0.93-1.05], .724	1.0 [0.95-1.05], .957	-	1.05 [0.98-1.11], .142	1.0 [0.95-1.06], .905
SITE 2	KNN	**1.0 [1.0-1.01], .038**	1.0 [0.91-1.11], .971	1.0 [0.89-1.05], .447	1.0 [1.0-1.01], .292	0.91 [0.81-1.03], .135	1.0 [0.9-1.07], .641	1.0 [1.0-1.0], .875	1.01 [0.9-1.13], .929	1.0 [0.92-1.1], .864
	LR	1.0 [1.0-1.0], .990	1.02 [0.93-1.12], .694	1.0 [0.9-1.05], .433	1.0 [1.0-1.0], .683	0.99 [0.9-1.09], .887	1.0 [0.94-1.09], .740	1.0 [1.0-1.0], .463	0.93 [0.83-1.05], .262	1.0 [0.94-1.14], .468
	SVM	1.0 [1.0-1.0], .680	**0.86 [0.76-0.97], .011**	1.0 [0.92-1.09], .889	1.0 [1.0-1.01], .212	0.96 [0.87-1.07], .475	1.0 [0.9-1.06], .535	1.0 [0.9-1.12], .949	1.0 [1.0-1.0], .775	1.0 [0.95-1.1], .673
	DL	1.0 [1.0-1.0], .956	0.95 [0.85-1.05], .289	1.0 [0.87-1.03], .192	1.0 [1.0-1.01], .275	1.04 [0.94-1.14], .485	1.0 [0.88-1.03], .241	1.0 [1.0-1.0], .344	0.96 [0.86-1.08], .518	1.0 [0.89-1.06], .503
SITE 3	KNN	1.0 [1.0-1.0], .522	0.95 [0.86-1.04], .253	1.0 [0.96-1.11], .456	1.0 [1.0-1.0], .916	1.06 [0.96-1.18], .236	1.0 [0.99-1.16], .092	1.0 [1.0-1.0], .741	0.96 [0.88-1.04], .268	1.0 [0.96-1.09], .515
	LR	1.0 [1.0-1.0], .516	1.01 [0.92-1.11], .799	1.0 [0.91-1.05], .561	1.0 [1.0-1.0], .285	0.97 [0.88-1.07], .548	**1.0 [1.0-1.2], .049**	1.0 [1.0-1.0], .098	1.0 [0.92-1.08], .942	1.0 [0.94-1.07], .930
	SVM	1.0 [1.0-1.0], .188	1.05 [0.95-1.15], .357	1.0 [0.89-1.04], .320	1.0 [1.0-1.0], .560	0.96 [0.87-1.05], .366	1.0 [0.95-1.12], .416	1.0 [1.0-1.0], .620	0.94 [0.82-1.08], .400	1.0 [0.89-1.1], .797
	DL	1.0 [1.0-1.0], .277	1.01 [0.92-1.11], .825	1.0 [0.91-1.06], .593	1.0 [1.0-1.0], .926	0.96 [0.88-1.06], .414	1.0 [0.96-1.11], .459	1.0 [1.0-1.0], .817	1.01 [0.93-1.09], .815	1.0 [0.94-1.08], .806
SITE 4	KNN	1.0 [1.0-1.0], .661	0.92 [0.79-1.09], .343	1.0 [0.78-1.02], .086	1.0 [1.0-1.0], .380	1.01 [0.92-1.12], .778	1.0 [0.93-1.11], .697	1.0 [1.0-1.0], .661	1.0 [1.0-1.0], .735	1.0 2 [0.91-1.12], .534
	LR	1.0 [1.0-1.0], .389	1.04 [0.94-1.15], .439	1.0 [0.84-1.0], .050	1.0 [1.0-1.0], .769	1.08 [0.99-1.19], .089	1.0 [0.89-1.04], .328	1.03 [0.95-1.1], .463	1.0 [1.0-1.0], .385	1.0 [1.0-1.0], .638
	SVM	1.0 [1.0-1.0], .584	1.01 [0.92-1.1], .894	**1.0 [0.85-0.99], .033**	1.0 [1.0-1.0], .126	1.04 [0.94-1.14], .482	1.0 [0.88-1.05], .355	1.0 [1.0-1.0], .884	1.0 [1.0-1.0], .472	0.96 [0.89-1.03], .264
	DL	1.0 [1.0-1.0], .621	1.04 [0.94-1.15], .493	1.0 [0.85-1.01], .072	1.0 [1.0-1.0], .572	0.95 [0.85-1.06], .356	**1.0 [0.82-0.99], .034**	1.0 [1.0-1.0], .521	1.02 [0.89-1.14], .573	1.0 [1.0-1.0], .647
SITE 5	KNN	1.0 [0.99-1.0], .207	0.99 [0.89-1.1], .847	1.0 [0.92-1.08], .947	1.0 [1.0-1.0], .375	1.04 [0.93-1.16], .464	1.0 [0.91-1.08], .828	1.0 [1.0-1.0], .489	0.96 [0.84-1.09], .535	1.0 [0.98-1.18], .128
	LR	1.0 [1.0-1.0], .195	**0.85 [0.75-0.97], .017**	1.0 [0.87-1.04], .267	1.0 [1.0-1.0], .306	1.06 [0.96-1.18], .260	1.0 [0.91-1.06], .615	1.0 [1.0-1.0], .475	0.94 [0.83-1.07], .365	1.0 [0.93-1.1], .836
	SVM	1.0 [1.0-1.0], .313	0.99 [0.89-1.1], .795	1.0 [0.85-1.02], .108	1.0 [0.99-1.0], .105	1.06 [0.94-1.18], .338	1.0 [0.86-1.02], .160	1.0 [1.0-1.0], .400	1.03 [0.91-1.16], .647	1.0 [0.87-1.03], .226
	DL	1.0 [0.99-1.0], .112	0.98 [0.87-1.1], .706	**1.0 [0.81-0.99], .027**	1.0 [1.0-1.0], .318	0.96 [0.86-1.06], .408	1.0 [0.98-1.15], .169	1.0 [1.0-1.01], .097	0.95 [0.83-1.08], .443	**1.0 [0.82-1.0], .048**

OR: odds ratio; CI: confidence interval; KNN: k-nearest neighbour; LR: logistic regression; SVM: support vector machine; DNN: deep neural network

3. eDiscussion

3.1. Association between sample size and classification accuracy

Accuracy and sample sizes from existing studies using ML and sMRI were extracted as follows:

- - up until and including 2013, this information was taken from the latest meta-analysis Kambeitz et al[31];
- - from 2014 to 2016 this information was taken from the review Arbabshirani et al[32;]
- - seven further subsequent studies were identified: Pinaya et al[33]; Salvador et al[34]; Winterburn et al[35]; Xiao et al[36]; Rozycki et al[37]; Dluhoš et al[38] and de Moura et al[39]. Xiao et al[36] was excluded as it was a clear outlier (see Figure 3a).
-

Pearson's correlation was used to test for the association between all sample sizes and accuracies. The same studies were used for Figure 1a.

3.2. Publication bias

Sample size, true positives, false positives, true negatives and false negative scores from each study as well as the overall main effect of ML-sMRI studies in psychosis (established schizophrenia and FEP combined) were extracted from Kambeitz et al[31]. Publication bias was assessed using the same procure as in Kambeitz et al[31] which in turn was based on recommendations for diagnostic classification studies described in Deeks et al[40]. Briefly, a measure of sample size and effect size were calculated as follows:

- Effective sample size (ESS) was calculated from the patients and control groups sample size using the formula:

$$1/\sqrt{ESS} = \frac{1}{\sqrt{\frac{4n1n2}{n1+n2}}}$$

- lnDOR (diagnostic odds ratio) was calculated using the following formula:

$$DOR = ln\left(\frac{TF/PN}{FP/TN}\right)$$

The resulting funnel plot was tested for asymmetry through a regression analysis weighted by ESS as implemented in R statistical programming language version 1.1.453 (R Core Team, 2016).

References

1. First MB, Gibbon M, Spitzer RL WJ. *Structured Clinical Interview for DSM-IV Axis II Personality Disorders*. American Psychiatric Press: Washington; 1997.
2. Bebbington P, Nayani T. The psychosis screening questionnaire. *Int J Methods Psychiatr Res*. 1995;5:11-19.
3. Andreasen NC, Flaum M, Arndt S. The Comprehensive Assessment of Symptoms and History (CASH). *Arch Gen Psychiatry*. 1992;49(8):615. doi:10.1001/archpsyc.1992.01820080023004
4. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage*. 2005;26(3):839-851. doi:10.1016/j.neuroimage.2005.02.018
5. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage*. 2007;38(1):95-113. doi:10.1016/j.neuroimage.2007.07.007
6. Yassa M, Stark C. A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *Neuroimage*. 2009;44(2):319-327. doi:10.1016/j.neuroimage.2008.09.016
7. Scarpazza C, Tognin S, Frisciata S, Sartori G, Mechelli A. False positive rates in Voxel-based Morphometry studies of the human brain: Should we be worried? *Neurosci Biobehav Rev*. 2015;52:49-55. doi:10.1016/j.neubiorev.2015.02.008
8. Mechelli A, Price C, Friston K, Ashburner J. Voxel-Based Morphometry of the Human Brain: Methods and Applications. *Curr Med Imaging Rev*. 2005;1(2):105-113. doi:10.2174/1573405054038726

9. Jones SE, Buchbinder BR, Aharon I. Three-dimensional mapping of cortical thickness using Laplace's equation. *Hum Brain Mapp*. 2000;11(1):12-32.

10. Hutton C, De Vita E, Ashburner J, Deichmann R, Turner R. Voxel-based cortical thickness measurements in MRI. *Neuroimage*. 2008;40(4):1701-1710. doi:10.1016/j.neuroimage.2008.01.027

11. Hutton C, Draganski B, Ashburner J, Weiskopf N. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *Neuroimage*. 2009;48(2):371-380. doi:10.1016/j.neuroimage.2009.06.043

12. Fischl B. FreeSurfer. *Neuroimage*. 2012;62(2):774-781. doi:10.1016/j.neuroimage.2012.01.021

13. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980. doi:10.1016/j.neuroimage.2006.01.021

14. Dale AM, Fischl B, Sereno MI. Cortical Surface-Based Analysis. *Neuroimage*. 1999;9(2):179-194. doi:10.1006/nimg.1998.0395

15. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 2002;33(3):341-355. doi:10.1016/S0896-6273(02)00569-X

16. Fischl B, Salat DH, van der Kouwe AJW, et al. Sequence-independent segmentation of magnetic resonance images. *Neuroimage*. 2004;23:S69-S84. doi:10.1016/j.neuroimage.2004.07.016

17. Jolliffe I. *Principal Component Analysis*. Springer, Berlin; 2002.

18. Lever J, Krzywinski M, Altman N. Points of Significance: Principal component analysis. *Nat Methods*. 2017;14(7):641-642. doi:10.1038/nmeth.4346

19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(Oct):2825-2830.

20. Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy. In:

Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016:308-318. doi:10.1145/2976749.2978318

21. Chollet F, others. Keras. 2015.
22. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am Stat*. 1992;46(3):175-185. doi:10.1080/00031305.1992.10475879
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol)*. 2005;67(2):301-320.
24. Pereira F, Mitchell T. Machine learning classifiers and fMRI : a tutorial overview. 2008:1-21. doi: 10.1016/j.neuroimage.2008.11.007
25. Vapnik V. *The Nature of Statistical Learning Theory*. Springer; 1995.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
27. Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev*. 2017;74:58-75. doi:10.1016/j.neubiorev.2017.01.002
28. Krogh A, Hertz JA. A Simple Weight Decay Can Improve Generalization. In: Lippman DS, Moody JE, Touretzky DS, eds. Advances in Neural Information Processing Systems, Vol. 4. Morgan Kaufmann; 1992:950-957.
29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
30. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010:249-256.
31. Kambeitz J, Kambeitz-Ilankovic L, Leucht S, et al. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern

recognition studies. *Neuropsychopharmacology*. 2015;40(7):1742-1751. doi:10.1038/npp.2015.22

32. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*. 2017;145:137-165. doi:10.1016/j.neuroimage.2016.02.079
33. Pinaya WHL, Gadelha A, Doyle OM, et al. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep*. 2016;6(38897). doi:10.1038/srep38897
34. Salvador R, Radua J, Canales-Rodríguez EJ, et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS One*. 2017;12(4):e0175683. doi:10.1371/journal.pone.0175683
35. Winterburn JL, Voineskos AN, Devenyi GA, et al. Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophr Res*. December 2017. doi:10.1016/ j.schres.2017.11.038
36. Xiao Y, Yan Z, Zhao Y, et al. Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophr Res*. December 2017. doi:10.1016/j.schres.2017.11.037
37. Rozycki M, Satterthwaite TD, Koutsouleris N, et al. Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. *Schizophr Bull*. 2018;44(5):1035-1044. doi:10.1093/schbul/sbx137
38. Dluhoš P, Schwarz D, Cahn W, et al. Multi-center Machine Learning in Imaging Psychiatry: A Meta-Model Approach. *Neuroimage*. 2017;155:10-24. doi:10.1016/j.neuroimage.2017.03.027
39. de Moura AM, Pinaya WHL, Gadelha A, et al. Investigating brain structural patterns in first episode psychosis and schizophrenia using MRI and a machine

learning approach. *Psychiatry Res Neuroimaging*. 2018;275:14-20. doi:10.1016/j.pscychresns.2018.03.003

40. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58(9):882-893. doi:10.1016/j.jclinepi.2005.01.016